usenix
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

acm In-Cooperation

SIGCHI

SOUPS 2014

Proceedings of the
Tenth Symposium On Usable Privacy and Security

Menlo Park, CA
July 9-11, 2014
http://cups.cs.cmu.edu/soups/

CyLab Usable Privacy & Security Laboratory
HTTP://CUPS.CS.CMU.EDU

Carnegie Mellon CyLab
CONFIDENCE FOR A NETWORKED WORLD

Proceedings sponsor:

commercenet

Foreword

The Tenth Symposium On Usable Privacy and Security featured 21 technical papers, three workshops, 29 posters, a panel, seven lightning talks, and an invited talk. We thank Facebook for hosting SOUPS 2014, which was held at their corporate headquarters in Menlo Park, CA.

This year we received 79 technical paper submissions. The program committee provided two rounds of reviews. In the first round papers received at least three reviews. This year, for the first time, authors had an opportunity to respond to the reviews their papers received in the first round. In the second round, papers that had received one or more reviews better than "weak reject" in the first round received additional reviews; in the end, papers received as many as six reviews. After a week of online discussion, the program committee held an in-person one-day meeting, which resulted in 21 papers selected for presentation and publication.

SOUPS 2014 featured an invited talk by Christopher Soghoian, Principal Technologist with the Speech, Privacy and Technology Project at the American Civil Liberties Union. Chris spoke about "Sharing the blame for the NSA's dragnet surveillance programs."

On Thursday evening SOUPS 2014 attendees enjoyed a dinner at Caffe Raice.The closing session on Friday featured a panel titled: "Division of labor between people and technology: just right or dumping the burden on users?" After a lively discussion, we concluded with the traditional SOUPS ice cream social.

This was the tenth year of SOUPS. We have grown from 71 attendees in our first year to a conference that sold out with close to 300 attendees registered.

We would like to thank all of the authors and the members of the technical papers committee and organizing committee for helping to produce this program. We are grateful to everyone whose assistance with logistical arrangements made this event possible: the staff at Facebook and the staff at CyLab and the Institute for Software Research at Carnegie Mellon University. We would like to thank CommerceNet for sponsoring the open-access publication of our proceedings by USENIX. We would also like to thank the US National Science Foundation, Facebook, Google, Blackberry, Cisco, Alston & Bird LLP, and CyLab for their sponsorship of this event, and USENIX for publishing our proceedings. SOUPS 2014 was held in cooperation with USENIX and ACM SIGCHI.

Lorrie Faith Cranor
General Chair
Carnegie Mellon University

Lujo Bauer
Technical Papers Co-Chair
Carnegie Mellon University

Robert Biddle
Technical Papers Co-Chair
Carleton University

ISBN 978-1-931971-13-3

SOUPS 2014 Organization

General Chair:	Lorrie Faith Cranor, Carnegie Mellon University, USA
Technical Papers Co-Chairs:	Lujo Bauer, Carnegie Mellon University, USA Robert Biddle, Carleton University, Canada
Invited Talks Chair:	Rob Reeder, Google, USA
Lightning Talks and Demos Chair:	Alain Forget, Carnegie Mellon University, USA
Local Activities Chair:	Maritza Johnson, Facebook, USA
Panels Chair:	Cormac Herley, Microsoft Research, USA
Posters Co-Chairs:	Mike Just, Glasgow Caledonian University, Scotland Yang Wang, Syracuse University, USA
Tutorials and Workshops Chair:	Sonia Chiasson, Carleton University, Canada

Technical Papers Committee:

Co-chair: Lujo Bauer, Carnegie Mellon University, USA
Co-chair: Robert Biddle, Carleton University, Canada
Konstantin Beznosov, University of British Columbia, Canada
Joseph Bonneau, Princeton University, USA
Sonia Chiasson, Carleton University, Canada
Sunny Consolvo, Google, USA
Alexander De Luca, University of Munich (LMU), Germany
Simson Garfinkel, Naval Postgraduate School, USA
Iulia Ion, Google, USA
Maritza Johnson, Facebook, USA
Apu Kapadia, Indiana University, USA
Wenke Lee, Georgia Tech, USA
Janne Lindqvist, Rutgers University, USA
Heather Lipford, UNC Charlotte, USA
Michael K. Reiter, UNC Chapel Hill, USA
Matthew Smith, University of Bonn, Germany
Melanie Volkamer, Technische Universität Darmstadt, Germany
Yang Wang, Syracuse University, USA
Tara Whalen, Carleton University, Canada
Mary Ellen Zurko, Cisco Systems, USA

Early-round External Reviewers:

Kelly Caine
Paul Dunphy
Serge Egelman
Alain Forget
Kirstie Hawkey
Patrick Gage Kelley
Brian LaMacchia
Michelle Mazurek
Emilee Rader
Rob Reeder
Elizabeth Stobert
Blase Ur
Rick Wash

Other External Reviewers:

Hala Assal
Natã Miccael Barbosa
Gradeigh D. Clark
Bryan Dosono
Xianyi Gao
Qatrunnada Ismail
Corey Jackson
Pooya Jaferian
Bart Knijnenburg
Jorge Guajardo Merchan
Xinru Page
Karen Renaud
Gerardo Reynaga
Sadegh Torabi
Na Wang
Huichuan Xia
Yulong Yang
Emanuel von Zezschwitz

SOUPS 2014 Awards

Distinguished Poster Awards

- *A Field Trial of Privacy Nudges for Facebook.* Yang Wang (Syracuse University); Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, Norman Sadeh (Carnegie Mellon University)
- *Computer security information in stories, news articles, and education documents.* Katie Hoban, Emilee Rader, Rick Wash, Kami Vaniea (Michigan State University)
- *Will this Onion Make You Cry? A Usability Study of Tor-enabled Mobile Apps.* Hala Assal, Sonia Chiasson (Carleton University)

IAPP Privacy Paper Award:

- *Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened thereafter? A study of the Westin categories, behavioral intentions, and consequences.* Allison Woodruff, Vasyl Pihur (Google); Alessandro Acquisti (Carnegie Mellon University); Sunny Consolvo, Lauren Schmidt (Google); Laura Brandimarte (Carnegie Mellon University)

Distinguished Paper Awards

- *Crowdsourcing Attacks on Biometric Systems*. Saurabh Panjwani (Independent Consultant) and *Achintya Prakash (University of Michigan)*
- *Understanding and Specifying Social Access Control Lists.* Mainack Mondal (MPI-SWS); Yabing Liu (Northeastern University); Bimal Viswanath, Krishna Gummadi (MPI-SWS); Alan Mislove (Northeastern University)

SOUPS Impact Award

On the 10 year anniversary of SOUPS, we have instituted a SOUPS Impact Award. The SOUPS Impact Award is to be presented every third year to the authors of a SOUPS conference paper that has had a significant impact on usable security and privacy research and practice. The winner of the award will be announced at the SOUPS conference.

To be eligible, a paper should have been published no earlier than 5 and no later than 10 years before the year in which the award is presented. For example, for the presentation at SOUPS 2014, the paper should have been published at SOUPS 2005 through SOUPS 2009.

The selection committee will focus on a paper's impact as judged by influence in the research community, a community of practice, industry, and/or social impact resulting from the work. A public citation for the award will be placed on the SOUPS website.

The initial selection pool is drawn from the top 10 most cited SOUPS papers of the eligibility period, excluding previous winners. The most highly cited paper need not be the winner, because citation counts are not necessarily accurate, and because citations are not the sole measure of impact. Instead, the winner will be selected by the Selection Committee after a vigorous online discussion.

The Awards Selection Committee shall consist of between 5 and 15 current or former program committee members. Members will be appointed by the selection committee chair(s). To be eligible, a committee member may not be an author or co-author of one of the initial selection pool. The committee co-chairs shall adjudicate conflicts of interest, appointing substitutes to the committee as necessary.

2014 SOUPS Impact Award

The following papers formed the initial selection pool for the SOUPS 2014 Impact Award, with their year of SOUPS publication, and the citation count that qualified them. Committee discussion narrowed the selection to a short list of finalists, from which one winner was selected.

- Reducing shoulder-surfing by using gaze-based password entry, M Kumar, T Garfinkel, D Boneh, T Winograd (2007, 124)
- A second look at the usability of click-based graphical passwords, S Chiasson, R Biddle, PC van Oorschot (2007, 135)
- **Finalist:** Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish, S Sheng, B Magnien, P Kumaraguru (2007, 144)
- Passpet: convenient password management and phishing protection, KP Yee, K Sitaker (2006, 151)
- **2014 Award:** Usability of CAPTCHAs or usability issues in CAPTCHA design, J Yan, AS El Ahmad (2008, 155)
- Web wallet: preventing phishing attacks by revealing user intentions, M Wu, RC Miller, G Little (2006, 157)
- Decision strategies and susceptibility to phishing, JS Downs, MB Holbrook, LF Cranor (2006, 168)
- **Finalist:** Password management strategies for online accounts, S Gaw, EW Felten (2006, 185)
- **Finalist:** Authentication using graphical passwords: effects of tolerance and image choice, S Wiedenbeck, J Waters, JC Birget, A Brodskiy (2005, 266)
- **Finalist:** The battle against phishing: Dynamic security skins, R Dhamija, JD Tygar (2005, 403)

Usability of CAPTCHAS Or usability issues in CAPTCHA design
by Jeff Yan and Ahmad El Ahmad
Published at SOUPS 2008

This paper discusses usability issues that should be considered and addressed in the design of CAPTCHAs. Some of these issues are intuitive, but some others have subtle implications for robustness or security. The paper proposed a simple but novel framework for examining CAPTCHA usability. The framework proposed was instrumental in promoting a structured approach to CAPTCHA usability and security. It pulled together emerging knowledge of practice, added some useful insights, presented a thoughtful analysis, and systematized it into the "goto" paper for people trying to design CAPTCHAs. The framework from this paper still holds today, and the paper continues to be cited in studies.

2014 Impact Award Committee Members:
Ross Anderson
Simson Garfinkel (co-chair)
Cormac Herley
Markus Jakobsson
Brian LaMacchia
Heather Lipford
Andrew Patrick
Diana Smetters
Mary Ellen Zurko (co-chair)

SOUPS 2014:
Tenth Symposium on Usable Privacy and Security
July 9–11, 2014
Menlo Park, CA

Thursday, July 10

Perspectives on Privacy

Warnings and Decisions

Users and Security

Friday, July 11

Mobile

Authentication

Social Networks and Access Control

Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences

Allison Woodruff
Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043
woodruff@acm.org

Vasyl Pihur
Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043
vpihur@google.com

Sunny Consolvo
Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043
sconsolvo@google.com

Lauren Schmidt
Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043
schmidtl@google.com

Laura Brandimarte
Carnegie Mellon University
5000 Forbes Av. HBH 2105C
Pittsburgh, PA 15213
lbrandim@andrew.cmu.edu

Alessandro Acquisti
Carnegie Mellon University
5000 Forbes Av. HBH 2105C
Pittsburgh, PA 15213
acquisti@andrew.cmu.edu

ABSTRACT

Westin's Privacy Segmentation Index has been widely used to measure privacy attitudes and categorize individuals into three privacy groups: fundamentalists, pragmatists, and unconcerned. Previous research has failed to establish a robust correlation between the Westin categories and actual or intended behaviors. Unexplored however is the connection between the Westin categories and individuals' responses to the *consequences* of privacy behaviors. We use a survey of 884 Amazon Mechanical Turk participants to investigate the relationship between the Westin Privacy Segmentation Index and attitudes and behavioral intentions for both privacy-sensitive scenarios and privacy-sensitive consequences. Our results indicate a lack of correlation between the Westin categories and behavioral intent, as well as a lack of correlation between the Westin categories and consequences. We discuss potential implications of this attitude-consequence gap.

1. INTRODUCTION

Privacy research pioneer Alan Westin conducted over thirty privacy-related surveys between 1978 and 2004 [25]. During this time, he developed a Privacy Segmentation Index consisting of three questions and a set of rules to translate participants' responses into three categories (fundamentalists, pragmatists, and unconcerned) [24, 25]. This index captures general privacy attitudes about consumer control, business, and laws and regulations. It has been hugely influential in the debate over privacy attitudes, and has been deployed by researchers in numerous studies, e.g., [13, 23, 26, 29].

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

Nonetheless, concerns have long existed regarding the predictive power of Westin's categories and the assumptions underlying his Privacy Segmentation Index. First, previous research has failed to establish a significant correlation between the Westin categories (which capture broad, generic privacy attitudes) and context-specific, privacy-related behaviors, either actual or intended [13, 23, 29]. Second, researchers have raised concerns regarding unstated assumptions underlying the index, which presumes individuals make privacy decisions that are highly rational, reflective, and informed [42]. Instead, scholars have posited that incomplete information or decision-making biases, among other factors, may cause a gap between the general attitudes captured by the Westin categories and actual, specific privacy behavior [4]. Third, the instrument has not been updated since approximately 1995, and it is not obvious that it remains current in our Internet-centric world.

It is perhaps unsurprising that generic attitudes (such as those captured by Westin's Privacy Segmentation Index) are poor predictors of context-specific behaviors [15]. The so-called privacy paradox is often interpreted as the apparent lack of correlation between privacy attitudes and behaviors, and much work on this topic has focused on contrasting generic attitudes with hypothetical or observed behavior. However, one might suppose that general attitudes would be more successful at predicting responses to consequences. For example, one might imagine that a fundamentalist would object more strongly than an unconcerned to a personal photo being distributed widely on the Internet. In this manuscript we test the relationship between the Westin categories and a diverse, large set of scenarios, and examine the previously unexplored connection between those categories and individuals' reactions to privacy-relevant outcomes from those scenarios. In other words, we examine whether generic privacy attitudes are correlated with individuals' attitudes and behavioral intentions when hypothetical but specific consequences arising from the protection or disclosure of personal information are described. We survey 884 Amazon Mechanical Turk participants to investigate this relationship.

Supporting but extending previous literature, our results suggest a lack of correlation between the Westin categories and any of the scenarios we designed, independent of the type of data, actions, and context presented to the participants. Expanding previous literature, our results also suggest a lack of correlation between the Westin categories and actual outcomes, regardless of the material consequences associated with the disclosure of personal data. We discuss the potential implications of this apparent attitude-consequence gap for the motives and rationales underlying privacy attitudes, and for the design and evaluation of privacy "personas" or privacy segmentations.

Additionally, we explore several potential improvements to the Westin Privacy Segmentation Index. First, we report on a data-driven segmentation of responses to the Westin questions, which did not result in significantly better response prediction than the Westin categories. Second, we explore the implications of making the Westin questions more specific by replacing generic companies with particular brands; our results indicate this manipulation tends to make participants less privacy-sensitive. Third, we investigate whether other specific variables such as personality traits and demographics are more predictive of responses to scenarios or outcomes than the Westin categories, and report that these variables have at best only slightly improved predictive power.

The rest of this paper is organized as follows. In the next section, we provide background information on the Westin Privacy Segmentation Index, as well as other related work. Next, we describe the methodology for our survey as well as describing supplementary data we gathered, and then we turn to findings. We next explore several potential improvements of the Westin Privacy Segmentation Index. We then discuss the implications of our work and conclude.

2. BACKGROUND

2.1 The Westin Privacy Segmentation Index

Beginning in the late 1970's, Westin conducted numerous privacy-related surveys, refining questions and category definitions over time [25]. In 1995, he introduced the Westin Privacy Segmentation Index (subsequently also called the Core Privacy Orientation Index), which he used for nearly a decade in order to make longitudinal comparisons [24, 25]. Note that the questions are specifically related to a consumer perspective, although they have been widely adopted in broader contexts, e.g., [13, 23]. This culminating set of questions is perhaps the most commonly known and used form of his survey instruments, and it is the one we have chosen to include in our study.

A survey using this index asks participants, "For each of the following statements, how strongly do you agree or disagree?" [1 = Strongly Disagree, 2 = Somewhat Disagree, 3 = Somewhat Agree, 4 = Strongly Agree]:

Q1: Consumers have lost all control over how personal information is collected and used by companies.

Q2: Most businesses handle the personal information they collect about consumers in a proper and confidential way.

Q3: Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today.

Based on their responses to these three questions, Westin used the following procedure for dividing participants into three categories [25]. First, responses to the individual questions are classified as follows:

> For Q1, responses of "Strongly Agree" or "Somewhat Agree" are considered privacy-concerned.
>
> For Q2 and Q3, responses of "Strongly Disagree" or "Somewhat Disagree" are considered privacy-concerned.

Next, participants are categorized according to the following rules:

1. Privacy Fundamentalists: Participants who give privacy-concerned responses to all questions;
2. Privacy Unconcerned: Participants who give responses that are *not* privacy-concerned to all questions;
3. Privacy Pragmatists: All other participants (i.e., participants who give a mix of privacy-concerned and not privacy-concerned responses).

In addition to these three questions, Westin drew on other items in his survey instrument to construct a representation of the categories. The essential meaning of these categories remained the same, although specific details varied over the years [25]. The 2002 Harris report provides the following representative descriptions of fundamentalists, pragmatists, and unconcerned [24]:

> **Privacy Fundamentalists**: At the maximum extreme of privacy concern, Privacy Fundamentalists are the most protective of their privacy. These consumers feel companies should not be able to acquire personal information for their organizational needs and think that individuals should be proactive in refusing to provide information. Privacy Fundamentalists also support stronger laws to safeguard an individual's privacy.
>
> **Privacy Pragmatists**: Privacy Pragmatists weigh the potential pros and cons of sharing information, and evaluate the protections that are in place and their trust in the company or organization. After this, they decide whether it makes sense for them to share their personal information.
>
> **Privacy Unconcerned**: These consumers are the least protective of their privacy – they feel that the benefits they may receive from companies after providing information far outweigh the potential abuses of this information. Further, they do not favor expanded regulation to protect privacy.

2.2 The Privacy Paradox

Numerous studies have documented an attitude-behavior dichotomy (also referred to as the Privacy Paradox), in which participants' privacy-related attitudes are seemingly at odds with their actual or intended behavior, e.g., [41, 4, 3]. Spiekermann et al. compared self-reported privacy preferences (as measured with an instrument building on Ackermann et al.'s work [1]) with actual disclosing behavior during an

online shopping episode, finding that participants did not live up to their self-reported privacy preferences [41]. Acquisti and Grossklags studied the relationship between general privacy attitudes and self-reported adoption of privacy preserving strategies and self-reported past release of personal information, and also found supporting evidence for the attitude-behavior dichotomy [4]. While, as noted above, it is unsurprising that general attitudes would not precisely predict context-specific behaviors [15], the dichotomy appears to apply not only to general attitudes and behavior but also to *specific* attitudes and behaviors: Acquisti and Gross demonstrated a gap between the information participants said they cared about protecting online, and what they were showing publicly on Facebook [3].

A number of studies have also documented an attitude-behavior dichotomy specifically for attitudes as established by the Westin categories, showing gaps between the Westin categories and actual behavior [29], the Westin categories and behavioral intentions [6, 20], and the Westin categories and specific attitudes [23]. Malheiros et al. reported that the Westin categories failed to predict disclosure of personal data items in an online setting [29]. Consolvo et al. reported that the Westin categories were not a good predictor of how participants would respond to requests for their location from social relations [13]. Jensen and Potts found inconsistent correlations between the decision to purchase in hypothetical e-commerce scenarios and the Westin categories (as established by an instrument they developed to classify participants into Westin categories) [20]. Further, in an investigation of California residents' attitudes toward law enforcement's access to cell phone location data, King and Hoofnagle found that the attitudes professed among fundamentalists, pragmatists, and the unconcerned did not align with Westin's descriptions of their attitudes [23].

Researchers have previously argued that the disconnect between general privacy attitudes (as measured by the Westin Privacy Segmentation Index or other instruments) and behaviors may be due to a multiplicity of non-mutually exclusive reasons. The reasons include: instruments such as the Westin Privacy Segmentation Index measure general attitudes, while behaviors are context-specific [15]; individuals may perform privacy calculus and make choices that are privacy-suboptimal because they are the most viable or convenient options, even if they are not in accordance with the individuals' privacy preferences [43, 44]; and/or individuals may lack awareness or information about privacy trade-offs, or be subject to various types of decision-making biases [2, 4]. Specific to the Westin Privacy Segmentation Index, King and Hoofnagle have proposed that the Westin categories and their predictive power may be weakening over time [23].

We build on this previous research on the attitude-behavior dichotomy by exploring the relationship between privacy attitudes (as measured by the Westin Privacy Segmentation Index), behavioral intent, and consequences. We believe this is a novel exploration of whether the attitude-behavior dichotomy extends to consequences.

Numerous studies have analyzed privacy concern, and applied diverse instruments for measuring it [34]. In addition to the Westin Privacy Segmentation Index, researchers have proposed other privacy scales, including the Internet Users' Information Privacy Concerns (IUIPC) scale [30] and the Privacy Concern Scale (PCS) [10], both of which contain more specific questions than the Westin Privacy Segmentation Index. Preibusch has observed that scenarios are one of the common ways of measuring privacy concern [34]. In focus group discussions with a small number of participants, Kwasny et al. introduced six brief scenarios relating to surveillance, location tracking, photo sharing, self-disclosure and relationship building, identity theft, and health disclosure [26]. Ackerman et al.'s work is one of the earliest to report the use of scenarios, and we have drawn on their work for inspiration as a representative example of this approach [1], adding outcomes to enable us to explore participants' responses to specific consequences. We believe this type of use of outcomes is novel and allows us to explore issues which have not been previously investigated, such as the relationship between attitudes and consequences as described above. We also believe we have explored a much wider range of scenarios than previously reported.

3. METHODOLOGY

We ran a two-phase study on Amazon's Mechanical Turk in January and February of 2014, which yielded complete data from 884 participants. We also conducted supplementary surveys on Google Consumer Surveys (GCS). In this section we provide details on our study goals, design, and administration, as well as information about the supplementary data and limitations.

3.1 Study Goals

Our study was broadly designed to explore the relationships among generic privacy attitudes (including the Westin Privacy Segmentation Index), responses to hypothetical scenarios, responses to outcomes, personality traits, and demographics. In this paper, we focus on the relationship between the Westin Privacy Segmentation Index and responses to hypothetical scenarios and outcomes.

Our interest in responses to hypothetical scenarios is not novel; Section 2.2 highlighted several studies that have used scenarios to capture individuals' context-specific privacy preferences. However, in this study, we test individuals' responses to a broader array of scenarios, covering diverse situations, types of data, and possible behaviors. In addition to that, we examine the relatively less explored connection between Westin categories and individuals' reactions to potential consequences arising from privacy-sensitive scenarios. In doing so, our goal was to examine whether, as we induce participants to consider a set of possible consequences of protecting or disclosing data (be those consequences negative or positive), individuals' generic privacy attitudes become relevant predictors of how an individual will subjectively perceive, or react to, those privacy trade-offs.

3.2 Study Design

We designed a two-phase study, which was reviewed and approved by CMU's IRB. Phase I consisted of a survey that included several measures of general privacy attitudes. We aimed to capture a wide range of concerns about online and/or offline contexts. After reviewing numerous scales, we chose four that best balanced the following criteria: frequency of use by other researchers, appropriateness for current online and offline environments, and differentiation from other scales that we included. Specifically, we included: the Westin Privacy Segmentation Index [24, 25]; the Westin Personal Privacy Question which is a single question "How concerned are you about threats to your personal privacy in

America today?" [Very Concerned, Somewhat Concerned, Not Very Concerned, or Not Concerned at All] that was used by Westin several times to measure broad public sentiment (it predates the Privacy Segmentation Index, but Westin continued to use it in at least one study after he introduced the Privacy Segmentation Index) [25]); the Internet Users' Information Privacy Concerns (IUIPC) scale which was introduced in 2004 by Malhotra et al. to measure online privacy concerns [30][1]; and the Privacy Concern Scale (PCS) which was introduced by Buchanan in 2007 to keep up with the changing world of online privacy and asks questions related to common online activities (registration, e-commerce, email) [10][2].

Phase I also included three questions which we designed to measure participants' degree of direct and/or indirect experience with misuse of personal information, drawing on questions such as those reported by Malhotra et al. for inspiration [30].

Finally, Phase I assessed personality characteristics using scales from the psychology literature. After carefully reviewing the literature and numerous personality scales, we chose the nine that best balanced the following criteria: relevance to privacy, prior validation, appropriateness for an online survey, and differentiation from other scales that we included. Specifically, we included: TIPI (Ten Item Personality Inventory) [17]; locus of control [35]; MFT (Moral Foundation Theory) [18]; general disclosiveness (subscales amount, depth and honesty) [19]; generalized self-efficacy [37]; SIRI (Stimulating-Instrumental Risk Inventory) [46]; ambiguity tolerance [28]; hyperbolic discounting [5]; and CRT (Cognitive Reflection Test) [16].

Phase II was administered to the same set of participants, and asked them to imagine themselves in three (out of 20) randomly chosen scenarios (see Appendix A for a complete list). Our focus was to compare general attitudes such as those captured by the Westin Privacy Segmentation Index to specific attitudes and behavioral intention when considering context-dependent scenarios and their respective outcomes. Hence, we created a set of privacy-relevant scenarios that manipulate the type of information participants were asked to imagine divulging or not divulging (financial, health, location, social, or otherwise), the context of the disclosures (for example, the party to whom the information was to be disclosed, online versus offline, whether or not the information was anonymized, when or if the information would be deleted) and the consequences of the disclosure (a range of positive and negative outcomes with different financial, health, social, and other impacts).[3] For example, Scenario 1 entertains the following situation: 'A marketing company offers you $1000 and free genetic testing in exchange for the rights to all your current and future medical records. They will have the right to resell or publish your data (anonymously or with information that could identify you, at their discretion)'.

The main response or dependent variable of this study was the answer to a question about likelihood of disclosure (henceforth **scenario response**). The specific question was "How likely would you be to [perform a given action]?" (on a 5-item Likert scale [1 = Not at all Likely, 2 = Slightly Likely, 3 = Moderately Likely, 4 = Very Likely, 5 = Extremely Likely]). For example, for Scenario 1, the exact wording was "How likely would you be to take the offer?"

We were also interested in additional variables that would allow us to better understand the participants' interpretation of and decision-making regarding the scenarios. Accordingly, in addition to the response variable measuring likelihood of disclosure, we also asked questions about participants' specific feelings about each scenario. Specifically, we asked about their confidence that they could make a good decision; how well they thought they could foresee what might happen if they disclosed the information; how risky they felt it would be to disclose the information; how much choice they felt they had about whether or not to disclose the information; how much control they thought they would have over what happened to the information if they disclosed it; how likely it was that they would be in this situation; and how advantageous/disadvantageous the scenario was overall, in the best case, and in the worst case for themselves, their friends and family, and members of society.

After participants responded to questions about three scenarios, we presented them with three outcomes for each scenario (randomly chosen from sets of scenario-specific outcomes) and asked them to make similar assessments in terms of attitudes and disclosure likelihood as they had originally done for the scenarios alone. For a given outcome, the **outcome response** is the participants' reported likelihood of agreeing to the scenario, assuming this was the only outcome. The outcomes represented a wide range of situations with positive, negative, or neutral implications for privacy or well-being. For example, one of the outcomes for Scenario 1 postulates that, 'Your medical data is combined with that of many others. It is used to find a new cure for a previously deadly disease. Neither you nor anyone in your family has this disease.' The complete text of the scenarios and outcomes appears in Appendix A. Based on cognitive testing in a pilot round, each participant was presented only three scenarios, plus three outcomes for each scenario, in order to minimize learning effects and fatigue. At the end of Phase II, demographic data was collected.

3.3 Survey Administration

We administered the survey on Amazon's Mechanical Turk (MTurk) platform in late January and early February of 2014.[4]

[1] We included three components of this scale, namely control, awareness, and collection. These are the novel components the authors introduced in [30]. The scale contains several additional components which are modifications of previous scales, some of which were originally designed for offline environments; we did not include these because they overlapped with other scales we included and/or because they appear less relevant in the contemporary context.

[2] We included the Privacy Attitudes component of this scale (with slight modifications to align the answer choices with other scales). We did not include the Privacy Behavior component which was less relevant for our purposes because of its focus on the use of specific technical capabilities.

[3] For this exploratory study, we did not manipulate the cross-product of all possible variables, but rather focused on the scenarios and outcomes that are most organic and natural based on a review of media reports, research reports, and our experience with participants' concerns in other studies. Future work would profitably include a more systematic manipulation of such variables.

[4] We also ran a pilot version of the survey in April of 2013, with a nearly identical survey instrument and approximately the same number of participants. We re-ran the survey in

For Phase I, MTurk workers were invited to complete a survey about personality and attitudes for a compensation of $2.50. Workers were required to have the following qualifications: live in the United States, Human Intelligence Task (HIT) approval rate >= 95%, and number of approved HITs >= 100. In the MTurk task description, we did not mention privacy to avoid biasing our population. The average completion time for Phase I was 18 minutes, making the average hourly compensation $8.20. This is roughly on par with the United States minimum wage and consistent with payment standards of the MTurk community. A total of 1000 workers completed the task for Phase 1. After data quality assessment, 27 turkers were removed from consideration due to failing catch questions and/or giving overly uniform answers to a large number of questions in a row. After allowing a week to pass in order to minimize potential priming effects from questions in Phase I, we invited the remaining 973 workers to complete Phase II for a compensation of $3.00. 884 individuals out of 973 (90.85%) recruited for Phase II completed it; data from all 884 of these participants is included in the analysis reported in this paper. The average completion time for Phase II was 17 minutes, making the average hourly compensation $10.66.

Table 1 shows several key self-reported demographic characteristics of this sample.

Table 1: Select demographic characteristics of the survey sample.

Demographic	Category	Frequency
Gender	male	47.07%
	female	40.39%
	other	0.31%
	prefer not to answer	0.21%
	skipped	12.02%
Age	18-24	19.84%
	25-34	38.85%
	35-44	15.01%
	45-54	8.02%
	55-64	5.34%
	65+	0.72%
	prefer not to answer	0.21%
	skipped	12.02%
Education	some HS	0.62%
	HS	9.15%
	some college	31.86%
	college	39.05%
	advanced degree	6.89%
	prefer not to answer	0.31%
	skipped	12.13%
Income in $	<20K	17.16%
	20-45K	29.29%
	45-70K	23.74%
	70-100K	9.56%
	>100K	5.65%
	prefer not to answer	2.57%
	skipped	12.02%

early 2014 (screening by MTurk ID to exclude prior participants) to ensure we had recent data to report, to correct a minor typo in Q3 (we also ran a GCS survey with and without the typo with 1500 participants in each condition and did not find a significant difference), and to test the robustness of the results across multiple administrations of the survey. Results from the pilot were largely similar to those reported in this manuscript, with the minor exceptions noted in Section 4.1, and are not included here for the sake of brevity.

3.4 Supplementary Data

We ran several supplementary studies on Google Consumer Surveys (GCS) to contextualize our analysis. These studies are not core contributions of this work, but are included as useful context for the reader. GCS is a market research tool that supports online surveys [22]. Internet users complete survey questions in order to access premium content, and publishers get paid as their users answer. Answers are anonymous and are not connected to personally identifiable information. Demographics (age, gender or geography) are inferred for some participants; this demographic information can be used to target questions to participants or to weigh the results.

In this paper, we include results from two GCS surveys. For both surveys, we targeted the general population in the United States, and we use raw data rather than weighted data for our analyses, as the inferred demographics may not be accurate [22].

First, we ran a GCS survey with the three questions from the Westin Privacy Segmentation Index with 1,500 participants in January 2014.

Second, we ran a GCS survey with 6,000 participants in February 2014 to explore participants' sensitivity to mentioning specific brands. It contained original and manipulated versions of the three questions from the Westin Privacy Segmentation Index, plus three additional questions about purchasing history and trust. This "Brand Survey" had six conditions (1000 participants per condition). In one condition, participants answered the original Westin questions. In the additional five conditions, participants answered the Westin questions modified to refer to Amazon, PayPal, Safeway, Visa, and Walmart rather than more generic terms such as "companies" or "businesses". After answering the three (modified) Westin questions, participants answered three questions about their frequency of past purchases at the specified company (or "online" for classic Westin), their intent to purchase from the specified company (or "online" for classic Westin) again in the future, and how trustworthy they found the company (consistent with Joinson et al's finding that there is a strong relationship between privacy and trust [21]). The full questions appear in Appendix B.

3.5 Limitations

The quality of responses and the composition of the sample are key issues in survey research. In this paper we focused on US respondents in order to reduce heterogeneity of the sample, and we leveraged MTurk and GCS. MTurk, which has been used in prior usable security and privacy research (e.g., [14, 8]), allowed us to collect data from a large number of diverse participants. Buhrmester et al. found that the MTurk population was significantly more diverse than typical American college samples and that using MTurk could result in data at least as reliable as that obtained using traditional methods [11]. Paolacci et al. similarly found evidence that MTurk yielded data comparable in quality to surveying on a university campus [33].

GCS also has limitations, for example its use of inferred demographics and the context in which questions are asked (brief surveys to access premium content) [22]. Nonetheless, some initial reports about GCS are encouraging. The Pew Research Center compared results for questions on a variety of subjects asked in telephone surveys to those obtained using GCS [22]. The median difference between results ob-

tained from Pew Research surveys and GCS was 3 percentage points, and the mean difference was 6 points. They also reported that the demographic profile of Internet users who respond to GCS is similar to that of Internet users in Pew Research Center surveys, and that technological use profiles are also fairly similar. A white paper from Google reports that GCS performed favorably against both a probability based Internet panel and a non-probability based Internet panel, based on several benchmarks [31]. As another example, New York Times' blogger and statistician Nate Silver reported that out of a wide selection of polls, GCS election polls ranked second in terms of accuracy and lack of bias in predicting the 2012 election results [39]. Further, Schnorf et al. administered a questionnaire with several identical privacy questions to multiple panels and report that the levels of privacy concern for both GCS and MTurk respondents were fairly similar to those of respondents in nationally representative samples [36].

Despite these encouraging findings regarding both MTurk and GCS, neither is likely to comprise a statistically representative sample of the general population. Callegaro et al. argue that not only do univariate statistics often vary across samples, but predictive relationships (including magnitude) can vary as well [12]. Future work would benefit from validation in a representative (or different) population, as well as investigation of cross-cultural issues.

Further, although hypothetical scenarios are often used for measuring privacy concern [34], clearly they do not directly measure behavior or attitudes. It would be valuable to extend our work by testing the predictivity of the Westin Privacy Segmentation Index for other indicators of privacy concern. Finally, as with all negative results, a definitive conclusion can not be drawn; our failure to find a correlation does not mean that none exists.

4. FINDINGS

In this section, we present data on the Westin Privacy Segmentation Index, and responses to scenarios and outcomes.

4.1 Westin Privacy Segmentation Index

Figure 1 shows the distribution of responses to the three Westin questions.[5] In Q1, agreement is privacy-concerned, while in Q2 and Q3, disagreement is privacy-concerned. Taking that into account, all three distributions have the same mode (the second-most concerned bucket).

Figure 2 shows the distribution of the Westin categories. Approximately 49% of participants are fundamentalists, 40% are pragmatists, and 10% are unconcerned.[6][7] For comparison, in Table 2 we include the distribution of Westin categories in several other surveys: the GCS survey we ran with only the Westin Privacy Segmentation Index (GCS1); the condition of the GCS Brand Survey we ran which began with the three unmodified questions from the Privacy Segmentation Index (GCS2); results from Westin's 2003 survey administered by Harris Interactive (we were not able to determine full details of this sample) [25]; and results from Westin's 2001 survey administered by Harris Interactive to 1529 members of the Harris Poll Online database, which were then weighted (although the details of the weighting are not fully provided for proprietary reasons) [24]. We provide these numbers so that the reader may better contextualize our results by making a qualitative comparison, but given the varying compositions of the samples it is difficult to draw any definitive conclusions. It appears to be the case that the MTurk population may contain more fundamentalists than the GCS population and the populations tested by Westin in 2003 and 2001. However, it is unclear whether this higher number is simply due to biases in the MTurk population, or whether it is in fact a more accurate representation of current national sentiment. Investigation with a nationally representative sample would be costly but informative.

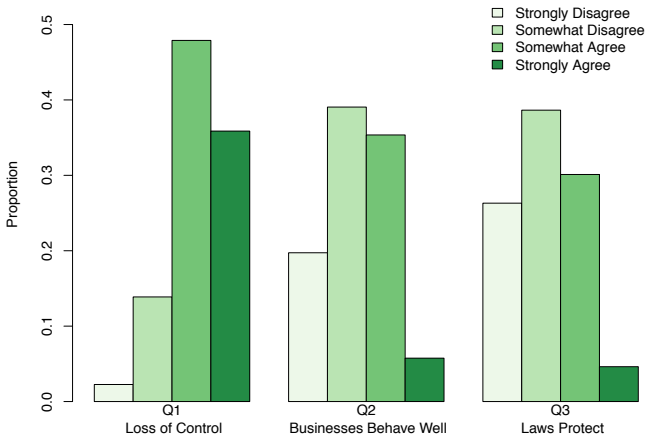


Figure 1: Distribution of raw scores for the three Westin questions. Note the mode of each distribution is the second-most concerned bucket.

We explored whether demographic variables predicted participants' Westin categories or their responses to the individual Westin questions. (Here and throughout, by 'predictive' we mean the ability to accurately predict the previously unobserved value of y based on the value of x given the observed relationship between the two variables in our

[5] For ease of reference we introduce brief précis for the three questions (e.g., 'Loss of Control' for Q1).

[6] Percentages do not sum to 100% due to missing responses.

[7] The alert reader may wonder if Snowden's revelations about NSA surveillance beginning in June 2013 affected the results [27]. Because we had conducted a pilot in April of 2013, we were able to compare data from before and after these events. There are marginally significant differences in responses to Westin's Q1 and Q3 before and after the NSA surveillance revelations. We found that both Q1 and Q3 showed increased concern of about 0.08 on the Likert scale after the NSA surveillance revelations, even after controlling for demographic differences. We did not find significant differences for Q2, nor did we find significant differences for the Westin categories (P-value: 0.8463 for the X^2 test). The minor shift in concern captured by Q1 and Q3 did not appear in the categorization because there was a shift to more extreme positions (from 'Somewhat Agree' to 'Strongly Agree' for Q1 and from 'Somewhat Disagree' to 'Strongly Disagree' for Q3) but not a change in polarity (the distribution of all 'Agree' answers and all 'Disagree' answers for a given question was relatively stable), and the Westin categorization rules rely on polarity. Overall, we found very few differences before and after the NSA surveillance revelations. For example, we found no changes in scenario response, with the exception of Scenario 13 about government surveillance of email, which participants were less likely to support post-revelation.

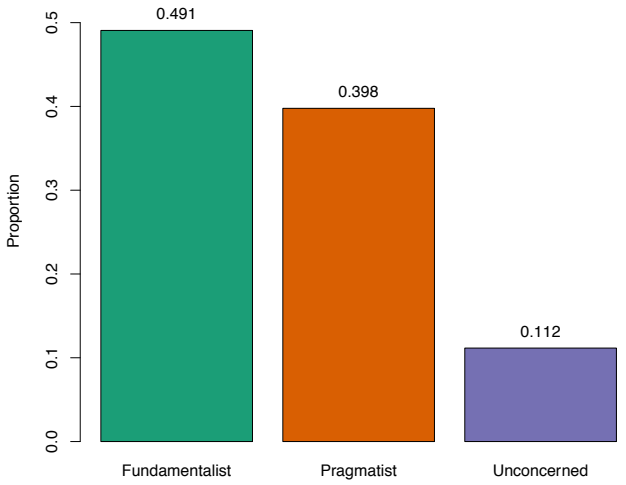


Figure 2: Distribution of the Westin categories.

Table 2: Distribution of the Westin categories in select data sets.

DataSet	Fundamentalist	Pragmatist	Unconcerned
MTurk '14	49%	40%	10%
GCS1 '14	38%	57%	6%
GCS2 '14	37%	58%	5%
Harris-Westin '03	26%	64%	10%
Harris-Westin '01	34%	58%	8%

sample.) Participants self-reported age, gender, education level, income, area where raised, area currently living, employment, religion and ethnicity. These demographic variables do not appear to be correlated with the Westin scale in our sample. No significant demographic predictors were found for any of the three individual Westin questions. We tested the association between the Westin categories and all the demographic variables using a X^2 test [40] with Monte-Carlo p-values because of the small counts in some table cells. Again, no significant associations were found.

We also explored whether any of the personality traits predicted participants' Westin categories or their responses to the individual Westin questions using separate one-way ANOVA models for each trait [40]. Full results are not shown due to space limitations, but in brief we found that purity, in-group, and authority (three dimensions of the MFT scale) have the highest predictive power for the three Westin categories, although the effects are modest (fundamentalists and unconcerned differ by at most 0.4 standard deviation units for any of the traits). The other three variables that show strong evidence of being correlated with the categories are locus of control, emotional stability and CRT; again the effects are modest. These six personality traits differentiated the fundamentalists from the pragmatists and unconcerned, although they revealed little differentiation between the latter two categories. These six traits had p-values < 0.00003 and were significant after correcting for multiple testing using the Bonferroni adjustment (0.05/76) [38]. Regarding the individual questions, similar results to the categories were found for Q2 and Q3, but not for Q1.

The reader will notice that we perform multiple tests for most of our analyses. Given the nature of this large exploratory study, it is critical to test a broad set of pre-defined hypotheses to narrow down the scope of studies that will follow. We are aware of the dangers of data snooping [45] and refrained from running additional analyses to discover 'interesting' results. In all cases, we used a Bonferroni correction to control the Type I error at the nominal level of 0.05 and to avoid an excessive number of false positive findings [38].

4.2 Scenarios

One of the main goals of this study was to examine the relationship between scenario response (i.e., likelihood of disclosure for a given scenario) and Westin categories. Each participant responded to three randomly chosen scenarios presented in a random order. The average sample size per scenario was 128 (min:109 and max:164). No significant differences were observed for any of the 20 scenario responses between Westin categories. Results are summarized in Figure 3. The x-axis lists the 20 scenarios and the y-axis shows scenario responses on a 5-point Likert scale, where 1 indicates 'Not at all Likely' to disclose and 5 indicates 'Extremely Likely' to disclose. Raw responses to scenarios are shown as colored dots (jittered) with three colors corresponding to the three Westin categories. Three solid colored lines trace the means for each category across the 20 scenarios. If Westin categories were significantly correlated with scenario responses, we would expect substantial divergence between the means lines. However, the data supports highly overlapping and crossing means and provides little evidence to the contrary. A formal analysis using one-way ANOVA models to test for differences in means between the three Westin categories for each scenario separately provides further evidence for the lack of association. Several marginally significant differences (Scenarios 3, 7, 11 and 13) disappear after the Bonferroni correction.

Proportions of variance explained by the ANOVA models, R^2's, range from 0% to 7% with a mean of 2% and give another indication of the insufficient ability of Westin categories to predict scenario responses. R^2 is a measure of the goodness of fit and is computed as a ratio of variance (in the response) that is attributed to the Westin categories divided by the total variance in the response.

Distributions of Westin categories within each response category are shown in the right margin of Figure 3. These are shown mainly for qualitative comparison to give the reader a sense of how the sizes of the three Westin categories differ for different response classes after combining all scenarios. If the Westin categories were predictive of the response, we would expect participants who answered 5 (Extremely Likely to disclose) to lean towards the unconcerned category, while those who answered 1 (Not at All Likely to disclose) would be mostly fundamentalists. We do not, however, observe substantial differences in terms of the distribution of Westin categories between these groups of participants.

The three Westin questions are framed in terms of consumer privacy, so we were also interested in comparing the predictive accuracy of the Westin categories for scenarios related to consumer privacy versus scenarios that did not have a consumer aspect. Two of the authors coded our 20 scenarios into three groups, with 100% agreement: three consumer-related scenarios (1, 3 and 4), six marginally consumer-

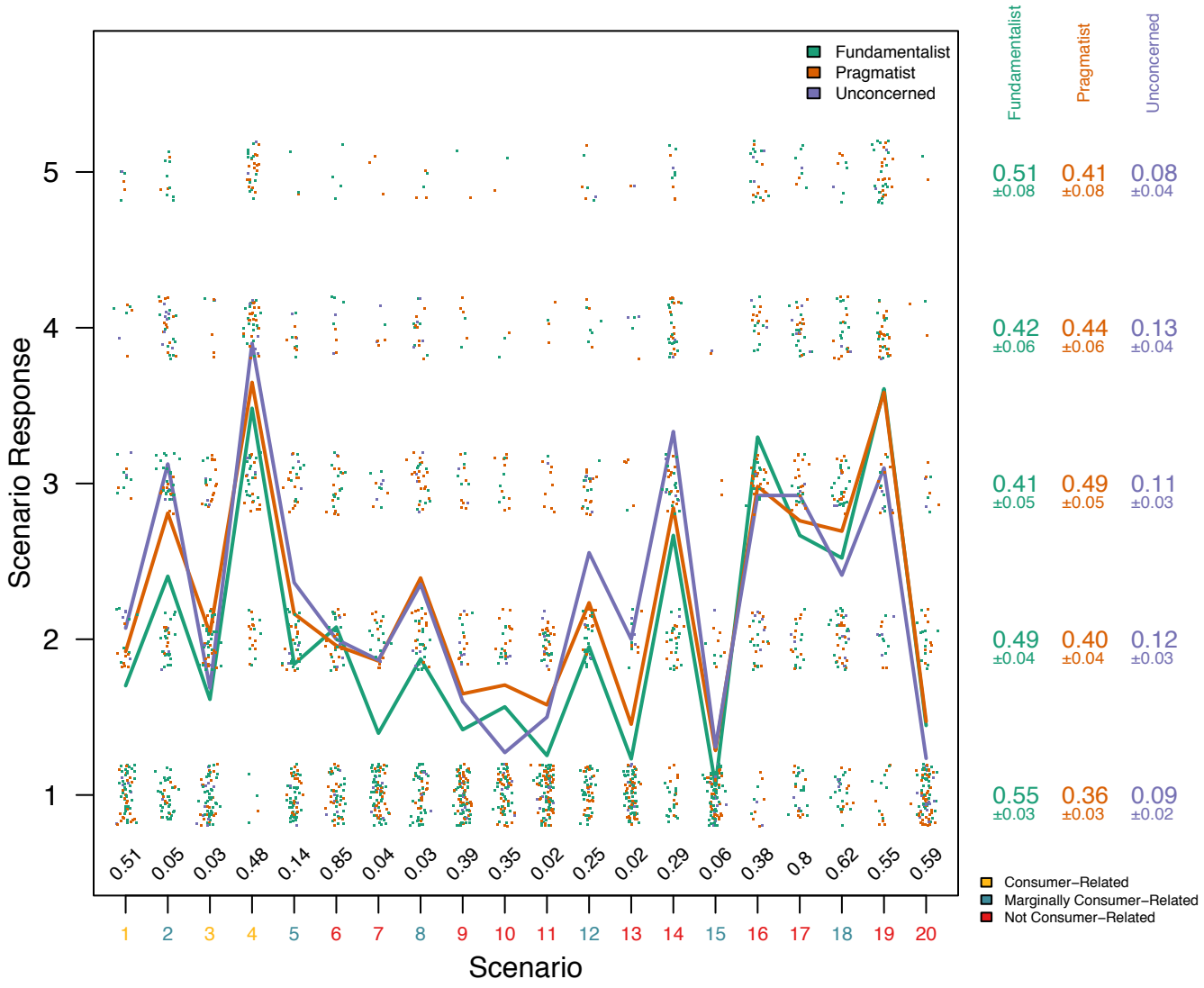


Figure 3: Only small differences (none significant with p-values shown at the bottom) were observed in the scenario response between the three Westin categories. Individual colored dots represent jittered scenario response with colored lines indicating the means for each segment. Scenario numbers at the bottom in different colors indicate the inferred scenario type and show no apparent patterns. In the right margin, the distribution of Westin clusters among each response category is shown with ± two standard deviations.

related scenarios (2, 5, 8, 12, 15 and 18), and 11 non-consumer related scenarios. These three types of consumer-relevance are shown in different colors in the labels for the x-axis of Figure 3. No clear difference emerges in terms of how Westin categories differ by consumer-relevance.

Just as the Westin categories are not predictive of the scenario response, individual Westin questions also show no significant associations (data not included for the sake of brevity). For all 20 scenarios, the proportion of variance explained by the three Westin questions ranges between 1% and 8%. We also note that the Westin categories do not appear to systematically predict any of the 15 scenario variables we collected. Only 6 of the 300 scenario-variable combinations (20 scenarios x 15 variables) had p-values < 0.001, and just 4 were significant after the Bonferroni adjustment.

4.3 Outcomes

In order to understand how increasingly specific information about situations affects responses, three randomly selected outcomes for each scenario were presented to participants. In total, 74 outcomes (3-5 per scenario) were considered and the average responses for each outcome for each Westin category are shown in Figure 4. The outcome response variable is the response to the question "How likely would you be to [disclosure specifics varied by scenario], knowing that this would be the only outcome?"

Clusters of three colored bars (one for each Westin category) represent an outcome. Scenario 1, for example, had

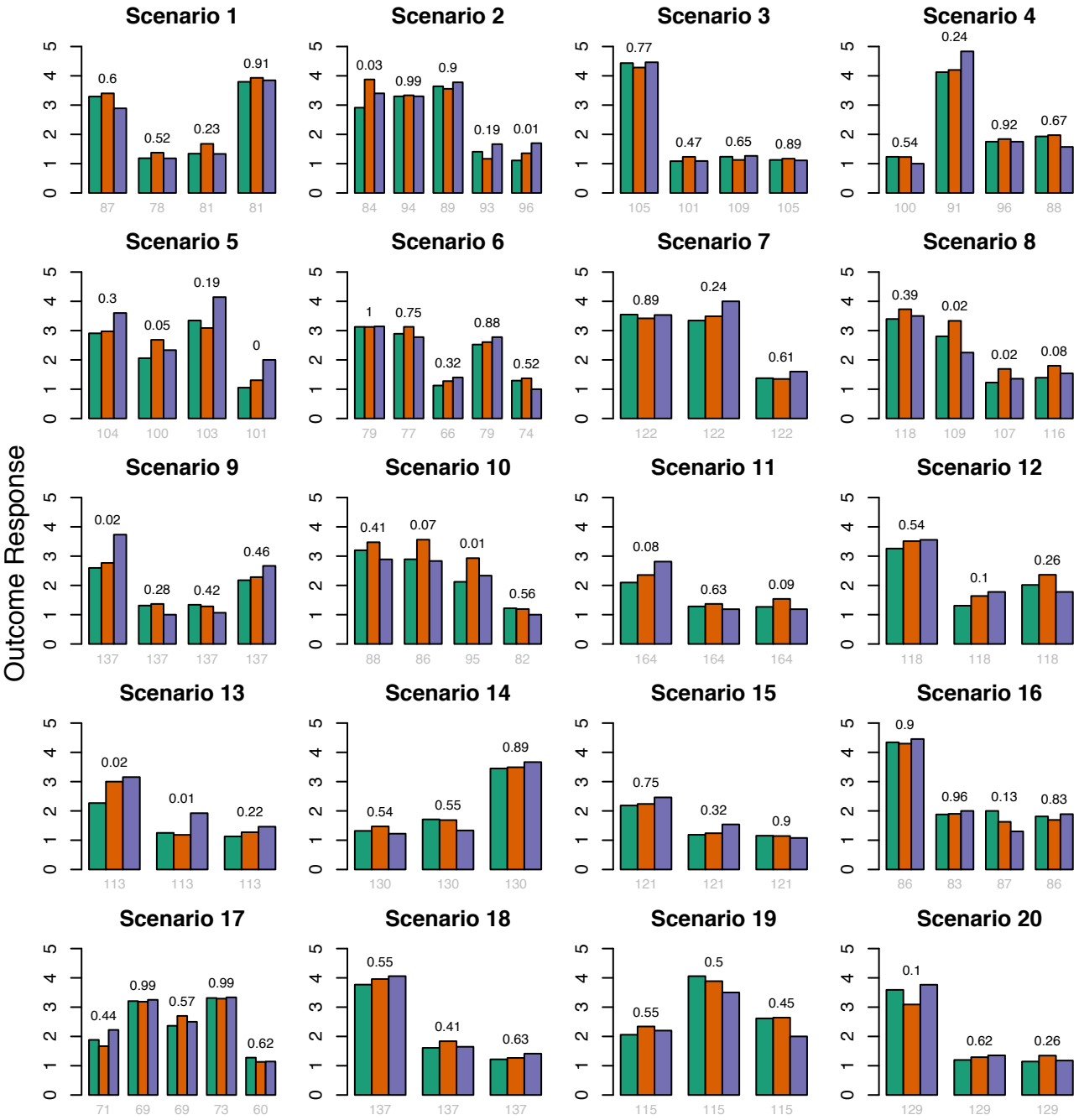


Figure 4: Westin categories by outcomes within each scenario are not significantly different for any of the 74 outcomes. P-values are shown at the top of each cluster of three bars, representing fundamentalists, pragmatists and unconcerned with the same colors as before.

four outcomes, while Scenario 2 had five outcomes. Counts in grey color under each combination of bars show the sample size of each outcome. If Westin's categories were significantly associated with outcome responses, we would observe bars of different colors having significantly different heights, but, as the figure shows, there is no systematic difference across the various outcomes of each scenario. P-values from a one-way ANOVA model [40] are shown at the top of each outcome cluster and indicate how different the Westin categories are in their response. Most outcomes do not show significant differences between the categories and none are significant after the Bonferroni correction. Please keep in mind again that with 74 tests, we would expect just under four of them to be significant prior to the Bonferroni correction even without any true differences. Overall, our results support the conclusion that Westin categories do not capture much of the variation present in the outcome response. As with scenarios, no significant differences were found for

the consumer-relevance of the outcome (p-value 0.5182).

Furthermore, individual Westin questions do not show significant associations with the outcome responses either. The total proportion of variance explained by the three individual Westin questions collectively ranges between 0.2% (Scenario 16, Outcome d) and 15% (Scenario 5, Outcome d) with a mean of 4.2%. Finally, although we do not present detailed data due to page limits, we note that the Westin categories also do not appear to predict any of the 5 outcome variables we collected. Only 5 out of 370 outcome-variable pairs (20 scenarios x 15 variables) had p-values < 0.001 and none were significant after the Bonferroni adjustment.

5. CAN THE WESTIN PRIVACY SEGMENTATION BE IMPROVED?

In the previous section, we failed to show a connection between the Westin Privacy Segmentation Index and responses to hypothetical scenarios and outcomes. In addition to the explanations that have been previously raised, in this section we explore three other possibilities. First, we explore whether different segmentation rules might yield a segmentation that is more predictive of responses to our hypothetical scenarios and outcomes. Second, we explore whether slightly modified versions of the Westin questions (made more specific by providing names of actual companies) yield different responses than the original questions. Third, we explore whether any of the other variables we measured were more predictive than the Westin Privacy Segmentation Index.

5.1 Data-Driven Segmentation

The Westin categories did not capture a significant amount of variation for responses to either scenarios or outcomes. However, it is possible that the three individual Westin questions capture more predictive information about individuals' privacy concerns but the segmentation rules themselves are not optimal and lead to an inferior separation ability.

To investigate this issue, we carried out a clustering of the Westin data using the k-means clustering algorithm with three clusters (other researchers have also used k-means clustering to classify subjects according to their privacy attitudes, e.g., [4, 41]). To visualize the relationship between how participants answer Westin questions and which group they are assigned to by the clustering algorithm, we present Figure 5. The three Westin questions are shown both in rows and columns. For example, the second panel in row 1 corresponds to Q2 in the x-axis and Q1 on the y-axis and shows the joint distribution of responses to these two questions. We invert the responses to Q2 and Q3 so that higher scores indicate more privacy concern. Each dot represents a pairwise response from a single participant, colored according to cluster. Responses are jittered to minimize overlap.

The three clusters found by the algorithm separate quite well, with clear clusters in the bottom left (least concerned, colored green), the middle (moderately concerned, colored blue), and the top right (most concerned, colored red) of each panel. The data-driven segmentation is somewhat different from the Westin one, and the distribution is different as well (cluster sizes are shown below the figure). Table 3 shows what happened to the original Westin categories during the new segmentation. The unconcerned group remains intact in Cluster 3 and receives an additional 52 participants from the pragmatist category. The pragmatist category loses

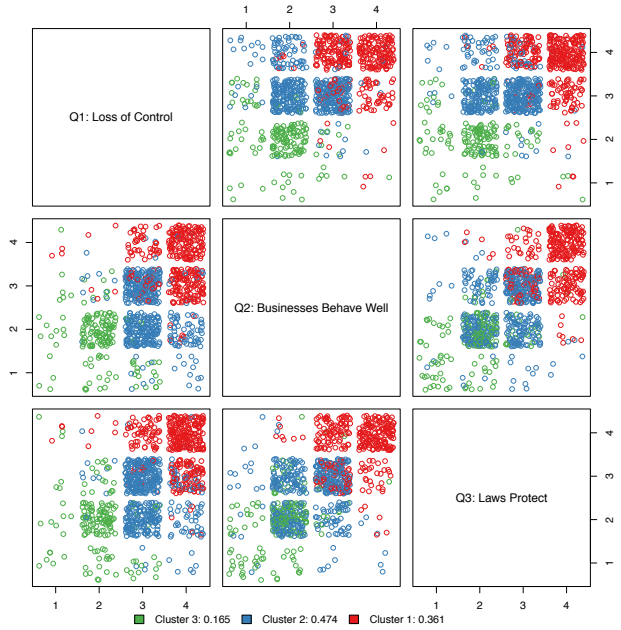


Figure 5: k-means clustering of Westin data with responses to Q2 and Q3 inverted so that higher scores on all questions indicate elevated concern. The x-axis and y-axis show the 1-4 Likert scale. Clear pairwise separation between clusters can be seen. Cluster sizes are shown below the figure.

Table 3: Data-driven segmentation in columns versus Westin categories in rows. A large portion of the difference is the split of the original fundamentalist category into Clusters 1 and 2.

	Cluster 1	Cluster 2	Cluster 3
Fundamentalist	329	146	0
Pragmatist	20	313	52
Unconcerned	0	0	108

an additional 20 participants to Cluster 1 (the new fundamentalist cluster). The largest difference between the two segmentations is the split of the original Westin fundamentalist category into two groups, which reduces the size of the new fundamentalist cluster significantly.

Because Westin prescribed specific segmentation rules, it is interesting to see what rules can be learned from the new segmentation. We use recursive partitioning [9] to that end (Figure 6). Q3 is the most informative of the three questions and is the first condition at the root of the tree. Thus, Q3 is the single variable that best splits the data into the two most homogeneous groups by maximizing the sum of the Gini index for the two nodes. The Gini index measures the impurity of the node in the tree and is defined as

$$1 - \sum p_i^2,$$

where p_i's are proportions of each class (in our case, proportions of each Westin category) in the node. After the initial split on Q3, the split on Q1 is critical for determining the

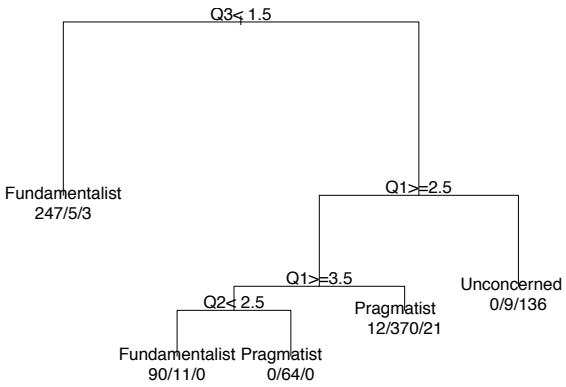


Figure 6: Data-driven segmentation rules for the three Westin questions. True conditions branch to the left and false to the right. The rules differ significantly from Westin's, with Q3 being the most important question to differentiate fundamentalists from others.

unconcerned, while Q2 picks up the remaining differences between fundamentalists and pragmatists. The learned rules for the new segmentation are as follows:

1. Privacy Fundamentalist: 'Strongly Disagree' on Q3 OR ('Strongly Agree' on Q1 and 'Strongly Disagree' or 'Somewhat Disagree' on Q2);

2. Privacy Unconcerned: Q3 is not 'Strongly Disagree' AND ('Strongly Disagree' or 'Somewhat Disagree' on Q1);

3. Privacy Pragmatist: All other participants.

Note that the learned rules do not perfectly reflect the new segmentation. The counts at the bottom in the format 'x/y/z' show how many participants from each cluster (fundamentalist/pragmatist/unconcerned) were classified into a particular category by that sequence of rules. For example, 5 pragmatists and 3 unconcerned answered 'Strongly Agree' on Q3 and are mistakenly attributed to the fundamentalist cluster.

We investigated how well the new segmentation predicts scenario and outcome responses. Results are practically analogous to the Westin categories, with the clusters showing little ability to differentiate participants' self-reported likelihood of disclosing. The clusters show the largest (yet still modest) separation of responses for Scenario 8, and this difference is statistically significant even after the Bonferroni correction (p-value 0.002). Similarly, Outcome b for Scenario 12 is also statistically significant after correction (p-value 0.0006). Overall, performing data-driven segmentation does not result in significantly better response prediction for either scenarios or outcomes.

5.2 Brand Manipulations

We wanted to investigate whether making the Westin questions more specific had an effect. As described above, we ran a GCS survey with 6000 participants in February 2014 to explore participants' sensitivity to mentioning specific brands (Amazon, PayPal, Safeway, Visa, and Walmart) rather than more generic terms such as "companies" or "businesses".

In fact, this small manipulation had a significant effect. Participants were significantly less concerned about privacy when considering a specific company. Table 4 shows differences by brand when compared to the original Westin questions. Brands are sorted by the largest difference in Q1. Very significant and practical differences appear between the five brands and the original general questions. Amazon, by all accounts, received the best marks, where Walmart and Visa yielded values closest to the original questions. Differences in individual questions translate into differences in Westin category frequencies. For the original Westin questions, we observed 37% fundamentalists, 58% pragmatists and 5% unconcerned. The proportion of fundamentalists was smaller for all brands (18% for Amazon and 34% for Walmart), with the proportion of unconcerned growing in all cases (to 25% for Amazon and 16% for Walmart). The trustworthiness variable had the largest effect on the Westin responses (results not shown), but did not explain away the significant differences between the brands after including it in the regression model (along with the other two measured variables about purchasing behavior).

5.3 What predicts disclosure?

We also explored whether personality traits, demographics, situational characteristics, or other privacy scales predicted either scenario or outcome response more effectively than the Westin Privacy Segmentation Index. To examine these relationships, we implemented a mixed-effect model using the *lme4 R* package [7]. Privacy attitudes, personality traits, demographics, and situational variables (all fixed effects) were regressed onto the scenario response along with two random effects (participant and scenario) to account for natural grouping in the data. Results were, perhaps, less encouraging than we hoped.

Analysis of the general privacy attitudinal scales (including the Westin Privacy Segmentation) revealed only small marginal effects that did not seem robust. However, four situational variables, namely, likelihood of the situation occurring, how advantageous the participant perceived the situation would be for them personally, how risky the situation was perceived to be, and how well the participant felt they could foresee the consequences of disclosing had the largest effects on the response. These effects and their corresponding 95% confidence intervals are 0.17 [0.124, 0.2], 0.15 [0.11, 0.195], -0.33 [-0.37, -0.29] and 0.06 [0.02, 0.1] on a Likert scale, respectively. Among the personality characteristics, only disclosure depth (effect size: 0.05 [0.002, 0.1]), disclosure amount (-0.074 [-0.13, -0.02]) and extraversion (0.05 [0.02, 0.09]) were statistically significant from 0. Here, effect size indicates by how much the response changes when the corresponding trait or characteristic changes by one unit. For example, considering the variable for the likelihood of the situation occurring, we would expect participants who answered 'Very Likely' to have, on average, 0.17 higher response scores than those who answered 'Somewhat Likely' given that every other variable remains fixed. The mixed model explains 59% of the response variance using the pseudo-R^2 measure developed as an analogue to the regular linear model. Of this 59%, 38% is attributable to fixed effects (variables we measured), and 21% is attributable to random effects (the participant and the scenario). No multiple testing correction was done here.

We performed the same analysis for the outcome response

Table 4: Differences in mean response for modified Westin questions by brand, as compared to mean response for the original Westin questions. The ± symbol indicates two standard deviations. Adjusted for age and gender.

	Visa	Walmart	Safeway	PayPal	Amazon
Q1: Loss of Control	-0.21 ± 0.081	-0.30 ± 0.080	-0.35 ± 0.081	-0.44 ± 0.080	-0.50 ± 0.080
Q2: Businesses Behave Well	0.12 ± 0.075	0.00 ± 0.075	0.07 ± 0.075	0.26 ± 0.075	0.33 ± 0.075
Q3: Laws Protect	0.20 ± 0.078	0.11 ± 0.078	0.18 ± 0.078	0.32 ± 0.078	0.40 ± 0.078

and obtained very similar results. The outcome mixed model included five additional outcome-specific variables and also the scenario-nested random outcome effect. This model explained about 61% of variance in the outcome response according to the pseudo-R^2 statistic. Of this 61%, 49% is attributable to fixed effects (variables we measured), and 12% is attributable to random effects (the participant and the outcome). Again, how risky the situation was perceived to be (effect: -0.1 [-0.13, -0.08]), how advantageous the participant perceived the situation would be for them (0.03 [0.003, 0.06]) and the likelihood of the situation occurring (0.05 [0.02, 0.08]) were significant scenario effects. All five outcome variables were also significant: how advantageous the participant perceived the situation would be for them personally (0.32 [0.29, 0.35]), how advantageous the participant perceived the situation would be for their friends and family (0.044 [0.01, 0.07]), how advantageous the participant perceived the situation would be for members of society (0.03 [0.006, 0.05]), the likelihood of the outcome occurring (0.18 [0.16, .2]) and how similar an outcome the participant imagined prior to viewing the outcomes (0.037 [0.02, 0.05]).

6. DISCUSSION

The Westin Privacy Segmentation Index is well-established, easy to administer, and yields design-relevant categories. However, consistent with but distinct from previous results, we failed to demonstrate a correlation between the Westin categories and either behavioral intentions or responses to consequences. While our failure to establish a correlation does not mean none exists, certainly the results are not encouraging. At this time, we can not recommend the use of the Westin categories to predict behavioral intentions or responses to consequences. Further, it may be wise to proceed with caution when deploying and interpreting results from the Westin Privacy Segmentation Index for other purposes, unless it has been established to be effective for them. Future work might productively explore whether alternative (e.g., [30, 10] or novel instruments (particularly those considering context [32]) have greater predictive power for both behavioral intentions and consequences.

While the lack of predictive power of Westin's categories across the hypothetical scenarios we presented to our participants is consistent with previous evidence of a gap between attitudes and behavioral intentions, our results also suggest a previously unreported dichotomy between attitudes and consequences. This lack of predictive power relative to actual outcomes can be interpreted in at least two different (and perhaps opposing) manners, suggesting the need for further research. One interpretation suggests that individuals reactions' are based on context-sensitive cost-benefit analyses (encompassing and mediated by complex factors such as systemic biases in decision-making) that are not captured by generic broad privacy attitudes. Another interpretation suggests that the Westin categories may instead capture some underlying, subjective, and deep-seated preferences for privacy that go beyond the so-called privacy calculus, and which may not be fully accounted for by the actual pros and cons of protecting or revealing data. We intend to investigate this further in future research.

A possible implication of these combined findings is that privacy segmentations, or privacy "personas," may inherently face ceilings in terms of their ability to predict privacy choices across diverse real life privacy conditions: there is an unavoidable trade-off between the clustering of preferences that privacy segmentations attempt to construct, and the specificity and heterogeneity of context-specific decisions. At the same time, said segmentations and personas may nevertheless help capture something deep and relevant about people's view of and preferences about privacy.

Finally, our scenarios and outcomes appear to be useful for studying participants' behavioral intentions and responses to consequences. We hope that this instrument may be useful to other researchers. For example, it might be used to explore the predictive value of novel segmentations, or to investigate whether an attitude-consequence gap appears for other instruments.

7. CONCLUSIONS

Previous research has established an attitude-behavior dichotomy, in which participants' broad privacy attitudes as measured by instruments such as the Westin Privacy Segmentation Index are seemingly at odds with their actual or intended privacy-related behaviors. However the relationship between attitudes as measured by the Westin Privacy Segmentation Index and specific consequences has not previously been explored in the literature. We conducted a survey to explore the relationship between the Westin Privacy Segmentation Index and participants' responses to a wide range of hypothetical scenarios and outcomes. We did not find evidence that either the individual questions or the derived categories of the Westin Privacy Segmentation Index are predictive of either participants' behavioral intent or their reaction to specific consequences, suggestive of both an attitude-behavior dichotomy and an attitude-consequence dichotomy. Future research might productively explore the inherent limitations of instruments for measuring broad privacy attitudes, while at the same time considering whether these attitudes capture underlying preferences that are not fully accounted for by contextual or practical considerations.

8. ACKNOWLEDGMENTS

We are grateful to Eyal Peer, Aaron Sedley, Jessica Staddon, and Joshua Tabak for inspiration and advice.

9. REFERENCES

[1] M. S. Ackerman, L. F. Cranor, and J. Reagle. Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 1–8. ACM, 1999.

[2] A. Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 21–29. ACM, 2004.

[3] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Privacy Enhancing Technologies*, pages 36–58. Springer, 2006.

[4] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 2:24–30, 2005.

[5] N. Ashraf, D. Karlan, and W. Yin. Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *The Quarterly Journal of Economics*, 121(2):635–672, 2006.

[6] N. F. Awad and M. Krishnan. The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, 30(1), 2006.

[7] D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2013. R package version 1.0-5.

[8] C. Bravo-Lillo, L. F. Cranor, J. Downs, S. Komanduri, and M. Sleeper. Improving computer security dialogs. In *Human-Computer Interaction–INTERACT 2011*, pages 18–35. Springer, 2011.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[10] T. Buchanan, C. Paine, A. N. Joinson, and U.-D. Reips. Development of measures of online privacy concern and protection for use on the Internet. *JASIST*, 58(2):157–165, 2007.

[11] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[12] M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, J. A. Krosnick, and P. J. Lavrakas, editors. *Online Panel Research: A Data Quality Perspective*. Wiley, 2014.

[13] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: why, when, & what people want to share. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 81–90. ACM, 2005.

[14] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. Does my password go up to eleven?: The impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2379–2388. ACM, 2013.

[15] M. Fishbein and I. Ajzen. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA, 1975.

[16] S. Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, pages 25–42, 2005.

[17] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37:504–528, 2003.

[18] J. Graham, J. Haidt, and B. A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96:1029–1046, 2009.

[19] B. Grams. Privacy concerns and personality traits influencing online behavior: A structural model. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 66(7-A):2421, 2006.

[20] C. Jensen, C. Potts, and C. Jensen. Privacy practices of Internet users: Self-reports versus observed behavior. *International Journal of Human-Computer Studies*, 63(1):203–227, 2005.

[21] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield. Privacy, trust, and self-disclosure online. *Human–Computer Interaction*, 25(1):1–24, 2010.

[22] S. Keeter and L. Christian. A comparison of results from surveys by the Pew Research Center and Google Consumer Surveys. *The Pew Research Center for the People and the Press*, November 2012.

[23] J. King and C. J. Hoofnagle. A supermajority of Californians support limits on law enforcement access to cell phone location information. *Social Science Research Network*, 2008.

[24] D. Krane, L. Light, and D. Gravitch. Privacy on and off the Internet: What consumers want. *Harris Interactive*, 2002.

[25] P. Kumaraguru and L. F. Cranor. Privacy indexes: A survey of Westin's studies. *ISRI Technical Report*, 2005.

[26] M. Kwasny, K. Caine, W. A. Rogers, and A. D. Fisk. Privacy and technology: Folk definitions and perspectives. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pages 3291–3296. ACM, 2008.

[27] S. Landau. Making sense from Snowden: What's significant in the NSA surveillance revelations. *IEEE Security and Privacy*, 11(4):54âĂŞ–63, 2013.

[28] A. MacDonald. Revised scale for ambiguity tolerance: Reliability and validity. *Psychological Reports*, 26:791–798, 1970.

[29] M. Malheiros, S. Preibusch, and M. Sasse. 'fairly truthful': The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure. In *Proceedings of the 6th International Conference on Trust & Trustworthy Computing (TRUST 2013)*, pages 250–266, 2013.

[30] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[31] P. McDonald, M. Mohebbi, and B. Slatkin. Comparing Google Consumer Surveys to existing probability and non-probability based Internet surveys. *Google White Paper*.

[32] A. Morton and M. A. Sasse. Privacy is a process, not a PET: A theory for effective privacy practice. In

Proceedings of the 2012 Workshop on New Security Paradigms, pages 87–104. ACM, 2012.

[33] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[34] S. Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. *International Journal of Human-Computer Studies*, 2013.

[35] J. Rotter. External and internal control. *Psychology Today*, 5(1):37–42, 1971.

[36] S. Schnorf, A. Sedley, M. Ortlieb, and A. Woodruff. A comparison of six sample providers regarding online privacy benchmarks. In *Proceedings of the Workshop on Privacy Personas and Segmentation at SOUPS 2014*, 2014.

[37] R. Schwarzer and M. Jerusalem. Generalized self-efficacy scale. *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*, pages 35–37, 1995.

[38] J. P. Shaffer. Multiple Hypothesis Testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

[39] N. Silver. Which polls fared best (and worst) in the 2012 presidential race. *The New York Times*, November 2012.

[40] M. Sirkin. *Statistics for the Social Sciences*. SAGE Publications, 1995.

[41] S. Spiekermann, J. Grossklags, and B. Berendt. E-privacy in 2nd generation E-commerce: Privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 38–47. ACM, 2001.

[42] J. Turow. Americans & online privacy: The system is broken. *Annenberg Public Policy Center, University of Pennsylvania*, 2003.

[43] J. Turow, L. Feldman, and K. Meltzer. Open to exploitation: America's shoppers online and offline. *Annenberg Public Policy Center, University of Pennsylvania*, 2005.

[44] J. Urban, C. Hoofnagle, and S. Li. Mobile phones and privacy. *UC Berkeley Public Law Research Paper*, 2012.

[45] S. S. Young and A. Karr. Deming, data and observational studies. *Significance*, 8(3):116–120, 2011.

[46] T. Zaleskiewicz. Beyond risk seeking and risk aversion: Personality and the dual nature of economic risk taking. *European Journal of Personality*, 15(S1):S105–S122, 2001.

APPENDIX

A. SCENARIOS AND OUTCOMES

1. A marketing company offers you $1000 and free genetic testing in exchange for the rights to all your current and future medical records. They will have the right to resell or publish your data (anonymously or with information that could identify you, at their discretion)
   (a) Your medical data is combined with that of many others. It is used to find a new cure for a previously deadly disease. Neither you nor anyone in your family has this disease.
   (b) Your data is published with information that identifies you. You lose a job due to your genetic information, which falsely suggests you may later develop a serious medical condition.
   (c) Your data is used to calculate the probability of certain diseases developing within your family. As a result, some of your relatives (but not you) see an increase of several hundred dollars a year in their health insurance premiums.
   (d) Your test results reveal that you have a serious but treatable disease of which you were previously unaware. You receive treatment just in time to make a full recovery.
2. You join an insurance plan which offers you the option of putting all of your health data in a unified healthcare database. All doctors, hospital staff, and emergency personnel will have access to these records without your needing to give any further permission
   (a) You avoid unnecessary duplicate vaccinations because your current doctor can see that you already received them.
   (b) You no longer have to fill out forms to transfer your medical records from one doctor to another.
   (c) Medical researchers combine your data with that of many other patients. The researchers notice geographic patterns and identify the outbreak of an epidemic much earlier than they would have otherwise. The outbreak, which is located far away from you or anyone you know personally, is contained before it spreads widely.
   (d) Marketers get access to the unified healthcare database and start sending advertising to patients being treated for addiction.
   (e) Your child's doctor looks up your health data and sees that you have been treated for depression. She alerts social services that they should look into whether or not you are caring well enough for your child.
3. Your friend tells you about a company that will give you free, customized investment advice. You go to the website, and to sign up you must provide detailed information about your income, credit history, investments, and investment goals.
   (a) You follow the investment advice and make a huge amount of money. You can quit your current job, retire, and travel the world.
   (b) The company sends you advice that is not helpful at all. They later use your information to commit credit card fraud in your name. They also attempt unsuccessfully to access funds in your bank accounts.
   (c) The company sends you advice that is not helpful at all, and sells your information to several banks. The banks use the information to predict the highest interest rates you personally are likely to pay, and send you targeted credit card and loan offers at precisely these rates. You accept one of the offers and end up paying higher interest than you would have otherwise.
   (d) The company sends you advice that is not helpful at all, and sells your information to several banks.

Based on the information you have provided about your investment goals, the banks conclude you are a poor credit risk and deny you a loan.

4. Your favorite retail store offers you a free loyalty card. You will save an estimated 10% on all store purchases you make when you present the card. To obtain the card, you are required to fill out a form with your name, address, and phone number, which may then be associated with a list of your purchases.
   (a) The retail store sells your data to your health insurance company. Your health insurance company analyzes your purchases, and concludes you have a sedentary lifestyle and an unhealthy diet. They raise your insurance rates.
   (b) You start receiving coupons from the retail store for products you frequently purchase. You end up saving 20% on your store purchases during the year.
   (c) Based on your purchasing patterns, the retail store builds a profile of you and sells it to national marketing companies. You receive tailored offers to which you are susceptible, and end up making some purchasing decisions you would not make normally and that you ultimately regret.
   (d) Your nosy neighbor works at the store. Against the company's rules, they look up the record of all your purchases. They learn that you bought some books about which you are slightly embarrassed. They tease you about the books, although they don't tell anyone else.
5. Your friends are all using a social networking application that lets them publicly share their location online, along with their first names. For example, whenever they arrive at a coffee shop or a bar, they can post that they are currently visiting that place. Your friends ask you to start using the application too, so you can coordinate social activities more easily.
   (a) You post that you are at your neighborhood coffee shop. Unbeknownst to you, a good friend is visiting from out of town. Your friend notices your post and stops by the coffee shop to say hi. You have a great time catching up.
   (b) You start receiving email coupons from the places you've visited, as well as shops near those places.
   (c) The editor at your city's newspaper notices that you go to a lot of performances by cool but obscure bands. They invite you to start writing music reviews for the paper, and you eventually become a minor celebrity.
   (d) A con artist looks up all the locations you have posted. They use the information to strike up a friendship with you, and they ask you for money for an investment opportunity. You invest several thousand dollars, and then you find out the investment opportunity was fraudulent. You feel betrayed and you never get your money back.
6. Your state starts offering a special GPS tag that you can attach to your car. If you have the tag, you can use a special fast lane whenever you go through a toll plaza, and your fare will automatically be charged to your account. Also, state and local agencies will be able to see everywhere you drive so they can manage traffic more effectively.
   (a) Traffic engineers study the GPS data from many users, and greatly improve traffic flow, public transit, and parking in your area.
   (b) Your city uses the GPS data to provide real-time traffic information, which saves you approximately 15 minutes of commute time per day.
   (c) The GPS technology reveals that you are speeding and you get a traffic ticket.
   (d) By using the fast lane at the toll plaza, you save approximately 5 minutes of commute time each day.
   (e) The database with drivers' full names and complete history of locations is hacked and made public. Information about a place you visit for personal reasons is revealed.
7. Your state starts offering a miniature digital monitoring device that can be implanted under a person's skin. The device monitors medical data such as heart activity and body temperature, and it also has GPS tracking to determine your location. The data can be accessed by government agencies and medical personnel in order to assist you or others, but it is not in a publicly available database.
   (a) You have an unexpected allergic reaction that requires immediate medical attention. The device detects the problem and alerts emergency medical personnel. They reach you in just a few minutes, and you make a full recovery.
   (b) You get lost while hiking in a remote area, but because you have the device, you are quickly found by rescue personnel and suffer no ill-effects.
   (c) The government compares GPS data from the devices with locations of crimes, in order to identify suspects. Based on your GPS data, you are wrongly accused of a violent crime and brought in for questioning, although you are quickly released.
8. You discover a free application for your cellphone that collects information about your activity and makes suggestions for improving your health. It automatically collects data on your exercise routes, speed, and duration; it lets you take pictures of food you are eating; it lets you track your sleep habits; and it occasionally asks you how you feel. It analyzes, graphs, and maps the data. It posts the data publicly online, without your name.
   (a) The application points out that you are more active and you feel better when you go to bed before 11pm. You change your habits to go to bed earlier every night. Because of this change, you reach your target weight, you are more productive at work, and you feel happier.
   (b) The application combines your data with that of many others in an anonymous way, and reveals that people feel worse when they go for a walk in your neighborhood. Scientists investigate and conclude that your neighborhood has high levels of pollutants from a local factory. The factory is shut down.
   (c) Someone at your workplace browses the publicly available data for people who live in the area. They

figure out which data is yours, and comment on the fact that you go for a walk every day in the park near their house.

(d) The application starts showing you targeted ads for businesses you pass on your daily walk.

9. Your city government proposes to install an extensive network of surveillance cameras and use face recognition technology to identify and track people as they move around the city.

   (a) The police quickly identify and capture a robber in your neighborhood.

   (b) The police arrest you for being in the proximity of a riot in which you did not participate.

   (c) You are turned away at the entrance to a sporting event because your face is very similar to that of someone who is banned from the stadium.

   (d) Election officials use the face recognition system to identify people who try to vote more than once. Because they eliminate this voter fraud, the candidate you are supporting for a local election wins.

10. The police department in your city proposes to purchase and deploy a fleet of small, low-flying unmanned aircraft that will fly around the city collecting audio and visual data. They explain that they can use this data for purposes such as monitoring city infrastructure or detecting unlawful activity, and they do not plan to make it publicly available.

    (a) The police department uses audio data to detect gunshots in a crime-ridden neighborhood, in which neither you nor anyone you know personally live. Because they are notified quickly when and where gunshots occur, the police are able to catch criminals and provide medical care to victims more efficiently. The crime in the neighborhood decreases quickly and numerous lives are saved.

    (b) During a bad storm, the video helps emergency personnel pinpoint key areas that are flooding. Because of this, they are able to build barricades that successfully protect people and property throughout the city that would otherwise have been injured or damaged.

    (c) The police use audio and video to monitor crowds during a large political protest. They are able to see where large numbers of people are building up and predict where riots are about to break out. They deploy additional security forces to these areas, thereby successfully quelling potential riots before they occur. No one is injured or arrested.

    (d) The police department computers are hacked. The hackers post all audio and video publicly online, including a recording of a very unpleasant fight you had with your significant other. Many of your friends and family see the recording and you feel embarrassed.

11. The political party you support wants to collect information about how individuals feel about various issues and candidates, in order to campaign more effectively. They create a website and ask their supporters, including you, to enter the names of people you know along with any information you have about their political leanings. For example, the website suggests that you enter the political orientation of your neighbors based on campaign signs you see displayed in their yards, and that you enter relevant information you glean from personal discussions with people you know.

    (a) The candidate you support wins the presidential election, in part because helpful volunteers such as yourself enter information about their friends and neighbors.

    (b) Your neighbor finds out you entered information about them. They are angry and cancel plans to have you over for dinner.

    (c) The political party sells the information to marketing companies. These companies use the information for targeted advertising, such as marketing guns to gun supporters and marketing liberal magazines to those who support gay rights.

12. A national newspaper starts publishing an online map that shows all political donations made by individuals. Anyone can search the map by name or address to see which causes an individual donated to, and how much. Many people start using it to look up donations made by people they know. You want to donate money to your favorite political candidate, but many of the people you know aren't aware that you support him.

    (a) Many of your friends see that you donated money, and they are inspired to donate to the candidate you support as well. The candidate you support wins the election, in part because of supporters like yourself and your friends.

    (b) Your boss finds out about your political leanings, and you are passed over for a promotion. You are pretty sure it is because your boss is unsympathetic to your beliefs, but you can't prove it.

    (c) Your next door neighbors find out about your political leanings. You hadn't realized it, but they strongly support the opposing party and they had assumed you did as well. Now every time you see them, they try to change your mind about how you are going to vote. They are polite but extremely annoying.

13. The government is considering passing a law to monitor all domestic email communications for security purposes.

    (a) The government identifies and averts a major terrorist attack.

    (b) The current administration analyzes many individuals' email to determine their political leanings. They use the information to redraw district boundaries and change hours at the polls, in order to give their political party an advantage in the next election.

    (c) Based on your email communications, you are wrongly accused and convicted of a crime you did not commit.

14. You cheated on your significant other, and you feel the need to talk to someone about it. You go out with your best friend, and sit in the corner of the bar. You believe no one can hear you. You consider whether or not to speak.

(a) Someone overhears you. Your significant other comes to know that you cheated on him/her, and breaks up with you.

(b) Your best friend reveals your secret to one of their friends that you are not very close to. However, your significant other never comes to know your secret.

(c) You feel better after discussing your secret with your friend. You recommit yourself to your relationship with your significant other.

15. You hear about a fun new game that you can play on your cell phone, and you think you would enjoy it. You learn that in order to play the game, you must enter the full names and email addresses of twenty of your friends.

(a) The gaming company sends email invitations to your friends to play the game with you. Several of them say yes. You enjoy playing the game occasionally, especially with your friends.

(b) The gaming company sells your friends' names and email addresses to a marketing company. Your friends start receiving annoying spam that appears as though it is from you.

(c) The game is a front for a scam. The gaming company sends email to your friends that looks like it is from you. The email says you are travelling internationally and are in trouble, and asks your friends to wire you money. Several of your friends fall for the scam and lose a total of several hundred dollars.

16. You move into a new home. One evening you overhear a loud fight at your next door neighbors' house. It sounds as though it might escalate into violence. You consider making an anonymous phone call to the police to report it.

(a) The police arrive promptly. You later learn that their presence probably prevented a violent episode, and the aggressor has now moved out of the house.

(b) The police arrive, but they don't find evidence of a problem and they leave. The loud fighting does not resume. However, one of your neighbors guesses you were the one who phoned and the next day they seek you out and tell you they are angry with you. You feel intimidated and you are worried they may retaliate against you in the future.

(c) The police arrive promptly. You later learn that the loud fight was actually the television and there was no problem. Your neighbor laughs it off.

(d) The police arrive promptly. You later learn that the loud fight was actually just a discussion about a football game that was on television. The police give your neighbors a ticket for disturbing the peace, and your neighbors have to pay a large fine.

17. You're at a party, and your friend makes a video recording of you doing a funny dance. They ask your permission to post it on a social networking site, saying they will only share it with mutual friends. You're sure it will make your friends laugh.

(a) Someone you've just started dating sees the video. They decide you are too silly for them and stop returning your calls.

(b) Your friends think it is awesome and compliment you on your moves.

(c) The person who posts the video gets the sharing settings wrong and anyone can see it. It goes viral and eventually appears in the mainstream media. You become a minor celebrity, known for being silly.

(d) One of your friends reshares it with a few people. An old friend from high school finds you because you're in the video. They get in touch with you and you're glad to hear from them.

(e) One of your friends reshares the video. It goes viral and is seen by a prospective employer. You don't get the job you were hoping for, because they think you are too silly to do well at the job.

18. You are planning a family vacation. Your friend recently had a great experience using a house swap website, and they recommend you try it. You go to the website and find a beautiful house in a terrific location in the city that you most want to visit. The owners have good reviews on the website from other people who have swapped houses with them in the past, and they are willing to swap houses with you for free at a time that is convenient for you.

(a) The house you visit is wonderful, and you have a great vacation. The family you swap with leaves your house in perfect condition.

(b) The house you visit is wonderful, and you have a great vacation. However, when you return home you can tell the other family riffled through all your things. Nothing seems to be missing, but you feel uncomfortable.

(c) The house you visit is messy and unpleasant, and you have a mediocre vacation. When you return home, you learn that the other family had a big party at your house and the police were called to break it up. The carpet is stained and several minor items are broken or damaged. You are unable to recoup the costs from the other family.

19. Your city is creating a time capsule that will be opened in 100 years. A photographer goes around town taking photos for the time capsule, and they take a photo of you and your significant other in a passionate embrace. They ask your permission to include the photo in the time capsule.

(a) The photo is included in a brochure describing the time capsule. The brochure is mailed to everyone currently living in your city. Your friends tease you and you are mildly embarrassed.

(b) No one sees the photo during your lifetime. When it is viewed in 100 years, it becomes an iconic image of romance in your time period, and you are immortalized.

(c) No one sees the photo during your lifetime. When it is viewed in 100 years, public displays of affection are frowned upon, and your behavior is considered scandalous.

20. You are searching for a job. You find an advertisement for a job that sounds perfect for you, and you start to complete the application online. When you're almost done, the form asks you to provide your social network

login and password so the human resources department can look at your private posts with friends and family. The form explains this will help the human resources department evaluate your fit with the company's culture.

(a) You are invited to interview, and you get the job. It is indeed a perfect fit for you. You make more money than you ever imagined, and you enjoy your work tremendously.

(b) You are invited to interview, and you get the job. However, you quickly discover the company engages in numerous unethical and illegal practices, and you resign before you get too embroiled in their wrongdoing.

(c) You do not get the job. However, one of the employees in the human resources department finds a private and moderately embarrassing photo of you, and posts it publicly on an Internet site that features such photos.

B. BRAND SURVEY QUESTIONS

In one condition, participants saw the original Westin Privacy Segmentation Index questions for the first three questions. In the other five conditions, participants saw the modified versions as specified below for the first three questions (modifications from the original questions are shown in bold font). Participants in all conditions saw the final three questions.

Consumers have lost all control over how personal information is collected and used by [**Amazon, Paypal, Safeway, Visa, Walmart**].

- Strongly Disagree
- Somewhat Disagree
- Somewhat Agree
- Strongly Agree

[**Amazon, Paypal, Safeway, Visa, Walmart**] handle**s** the personal information **it** collect**s** about consumers in a proper and confidential way.

- Strongly Disagree
- Somewhat Disagree
- Somewhat Agree
- Strongly Agree

Existing laws and organizational practices provide a reasonable level of protection for [**Amazon, Paypal, Safeway, Visa, Walmart**] consumer**s'** privacy today.

- Strongly Disagree
- Somewhat Disagree
- Somewhat Agree
- Strongly Agree

How many times have you made a purchase [from Amazon, with Paypal, from Safeway, with Visa, from Walmart, online] within the past 12 months?

- Never
- 1 time
- 2 - 5 times
- 6 - 10 times
- More than 10 times

How likely is it that you will make a purchase [from Amazon, with Paypal, from Safeway, with Visa, from Walmart, online] within the next 12 months?

- Not at all Likely
- Slightly Likely
- Moderately Likely
- Very Likely
- Extremely Likely

How trustworthy [is/are] [Amazon, Paypal, Safeway, Visa, Walmart, online vendors]?

- Not at all Trustworthy
- Slightly Trustworthy
- Moderately Trustworthy
- Very Trustworthy
- Extremely Trustworthy

Parents' and Teens' Perspectives on Privacy In a Technology-Filled World

Lorrie Faith Cranor, Adam L. Durity, Abigail Marsh, Blase Ur
Carnegie Mellon University
{lorrie, adurity, acmarsh, bur}@cmu.edu

ABSTRACT

The life of a teenager today is far different than in past decades. Through semi-structured interviews with 10 teenagers and 10 parents of teenagers, we investigate parent-teen privacy decision making in these uncharted waters. Parents and teens generally agreed that teens had a need for some degree of privacy from their parents and that respecting teens' privacy demonstrated trust and fostered independence. We explored the boundaries of teen privacy in both the physical and digital worlds. While parents commonly felt none of their children's possessions should ethically be exempt from parental monitoring, teens felt strongly that cell phones, particularly text messages, were private. Parents discussed struggling to keep up with new technologies and to understand teens' technology-mediated socializing. While most parents said they thought similarly about privacy in the physical and digital worlds, half of teens said they thought about these concepts differently. We present cases where parents made privacy decisions using false analogies with the physical world or outdated assumptions. We also highlight directions for more usable digital parenting tools.

1. INTRODUCTION

In the last twenty-five years, the daily life of a teenager has changed drastically. When the parents of today's teenagers were themselves teens, they had no smartphones connecting them to resources across the globe in an instant. In fact, except in rare cases, they had no mobile phones at all. Twenty-five years ago, teenagers only had access to the Internet at college or via Prodigy, Compuserve, or AOL. Stanley Milgram was the king of social networks; Mark Zuckerberg was just starting elementary school. Photos were developed in a darkroom or on Polaroid film, not Snapchatted.

While parenting has always been tough, these rapid shifts in technology create additional challenges for today's parents. Teenagers are more likely than their parents to understand popular technologies, services, and devices. They are also likely to socialize with friends using these technology-mediated channels. As a result, parents cannot necessarily draw from their own teenage experiences when making decisions about privacy for their children.

In this paper, we investigate how parents make decisions about privacy for their teens in a world that is far different than the one in which they came of age. We focus on parents' privacy decision making, as well as both teens' and parents' perspectives on the degree to which teenagers should have privacy from their parents. Through interviews, we explored four main research questions about teen privacy:

1. From teens' and parents' perspectives, what are the bounds of teens' right to privacy from their parents?
2. How do parents decide how much privacy teens should have when they use new technologies and services?
3. How do parents use parental controls, monitoring software, and ad-hoc approaches regarding teen privacy?
4. How do parents' approaches to privacy in the digital world compare to those in the physical world?

To investigate these research questions, we conducted semi-structured interviews with ten teenagers and ten parents of teenagers. Interviews covered teen privacy in the familiar physical world (e.g., closed doors and dating), in the technology-mediated digital world (e.g., smartphones and social media), and from a philosophical perspective. We focused our questions and analysis on privacy in the digital world and on parents' decision making process, using privacy in the physical world and on a philosophical level to contextualize attitudes about privacy in the digital world.

We found that most of our parent and teen participants agreed that teens should have privacy from their parents, albeit to a limited extent. This right to privacy derived from factors like trust and the desire to foster independence, but was limited by reasons including parental concern and safety. In the physical world, parents generally gave teens some degree of private space at home and in their social lives, such as by knocking before entering a bedroom.

In contrast to parents, teenagers viewed their cellphones, especially text messages stored on their cell phones, to be particularly private. Eight of the ten teens, versus four of the ten parents, felt it unethical for parents to look through teens' text messages. Teens were far more comfortable with their parents accessing their email accounts or Facebook, both of which they used rarely.

We unpack parents' processes for evaluating and regulating their children's privacy, finding that parents largely struggle to make these decisions. In particular, our parent

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

participants often did not understand teens' use of technologies that did not exist when the parents were themselves teens. While half of the teen participants said they think about privacy in the digital world differently than in the physical world, only two parents distinguished between these scenarios. Even though most parents wanted to give their teens private space, they did not always realize the degree to which teens' private spaces are text messages and apps.

Our results aid in understanding the complex issue of privacy as teenagers transition from dependent children to independent adults. This understanding can inform designers of software tools that directly or indirectly impact teen privacy. We discuss the shortcomings of existing digital parenting tools; we also speculate on directions for designing tools that better remind teenagers of their parents' expectations and help parents navigate the complex process of making decisions about their children's privacy.

2. BACKGROUND

Privacy is a complex concept that means different things to different people. Over a century ago, Warren and Brandeis discussed privacy as the "right to be let alone" [30]. In more modern interpretations, Helen Nissenbaum explained privacy through the idea of contextual integrity [16], while Daniel Solove proposed that privacy is best examined as a family of related concepts [22] and that privacy can be both an individual and a societal good [23].

Privacy as a legal right is even more complex. Privacy laws in the United States are sectoral, varying by industry. While several amendments within the U.S. Bill of Rights have been interpreted as providing some baseline privacy protections to United States citizens [24], most U.S. privacy laws are enacted to address specific concerns. The state of privacy protection in practice, however, often differs from the laws on the books [1]. In many cases, individuals can have de facto rights through social norms and beliefs. This difference between legal definitions and practice is particularly relevant to teenagers because teenagers have few legal rights to privacy from their parents. In practice, however, many parents do give teenagers some degree of privacy.

While the scholarship on privacy rights and laws is broad, it tends to focus on intrusions on individuals' privacy by the government or corporations. Far less has been written about privacy between individuals, and more specifically the aspect of privacy examined in this paper: the privacy beliefs and expectations between parent and teenager, especially in regards to digital space. Researchers who study teenagers and technology [12, 18, 32] have identified a surprising lack of studies investigating the role privacy plays in parent-teen relationships, and vice versa.

Marwick et al. surveyed the literature on youth and privacy [12]. They note the particular importance of studying teen privacy relative to technology since much of teens' socialization is mediated by technology. They also highlight findings that teens care deeply about privacy, particularly from parents and teachers [12]. boyd's recent book synthesizing years of fieldwork discusses the privacy dynamic between teenagers and parents [2]. She found that teenagers are quite concerned about having privacy from their parents and that parents rarely grant teens privacy without teens negotiating for it. She asserts that even well-intentioned parents "often...fail to realize how surveillance is a form of oppression." Whereas teenagers are boyd's primary subjects, we split our investigation equally between parents and teens. Recently, Ur et al. [27] interviewed both teens and parents in the more narrow context of home security systems with audit logs. They found that such Internet-connected home technologies have the potential to harm teen privacy while at the same time improving home security.

Researchers have also investigated teen privacy from parents' perspective. In her book on modern parenting, Nelson notes the hypocrisy of some parents in monitoring their teenagers while at the same time stating that they believe teens have a right to privacy [15]. Petronio describes ways in which parents invade their children's privacy, as well as teens' reactions ("children's defensive behaviors") to these invasions of privacy [17]. She notes that parents' and teens' divergent expectations of independence may cause conflict, yet did not explore the gap between parents' and teens' perspectives on teen privacy in any detail. Hawk et al. also found teenagers' perceptions of parental privacy invasion causes conflict, yet the magnitude of conflict differs by family [8]. However, they note that parent-teen conflict sometimes plays a positive role in adjusting parents' expectations.

Synthesizing recent psychology research, Smetana et al. points out that parents often adjust their parenting styles and attitudes for their different children [21]. Yardi and Bruckman also note that a particular child's maturity is a major factor in parents' decisions [32]. In separate work, Smetana found that parents generally reduce the extent to which they monitor their children as the children progress through adolescence [21].

While some parents monitor their teens closely, other parents prefer not to monitor teens at all [32]. Rode conducted in-home studies of twelve households with children, identifying five major strategies parents use to enforce rules about technology [18]. Some participants actively chose to use software tools to monitor teens' activities, while others preferred to talk with their children about safe behaviors. Metzger et al. explain that parents' differing opinions of teens' right to privacy, as well as the trust in the parent-teen relationship, led the parents they studied to reach different conclusions about the ethics of parental monitoring [14].

Parental monitoring sometimes leads children to share less information with parents. This phenomenon has been documented in traditional disclosure settings, such as conversation [8], but also on social media [9]. Teens' voluntary disclosure of information depends heavily on having a positive relationship with their parents [20]. Livingstone and Bober argue that strict monitoring can "undermine the democratic negotiation of mutual rights, trust and responsibilities between parents and children" [10].

Researchers have also investigated the adoption, as well as the non-adoption, of technologies parents can use to monitor their children. Vasalou et al. conducted a survey of 920 parents to understand why so few parents adopt technologies for tracking their children's location [28]. Many of their participants felt such systems could negatively impact their children's independence, suggesting that parents did feel that children have a right to privacy from their parents. Reaching similar conclusions, Czeskis et al. applied Value Sensitive Design to a series of parental monitoring scenarios, suggesting that context be taken into account when parents are deciding whether or not to monitor their teens [6]. Based on their analysis, they suggested that teens could have privacy from their parents except in the case of emergency.

In recent years, social media sites have become a battleground for teenagers' privacy from their parents. Although 80% of parents who used social media had friended their teenager on a social media site, many teens were uncomfortable with this practice [11]. Child and Westermann conducted a 235-participant survey related to parental Facebook friend requests, finding that teenagers generally accepted these requests out of obligation without making substantial changes to privacy settings [5]. In contrast, Cheng found that teens use a number of creative strategies to protect their online privacy, such as temporarily deactivating their Facebook account except when they decide to log in [4]. Forte et al. found that many high school students self-censor and maintain different social networks on different sites [7].

Teens are often more tech-savvy than their parents, leading parents to feel outmatched when attempting to monitor their children [25]. Wisniewski et al. conducted semi-structured interviews with ten pairs of parents and teens. Among their findings was that parental ignorance of technology could impede the parent's ability to engage meaningfully in the teen's online activity [31].

In contrast to much of this past work investigating teen privacy, we adopt a structured methodology that equally investigates the perspectives of teenagers and parents of teenagers. We rely on both perspectives to understand how parents make decisions about their children's privacy in a world that is far different from the one in which they came of age. We also document the extent to which our participants believe that teens have a de facto right to privacy from their parents in the absence of legal rights to privacy.

3. METHODOLOGY

We conducted semi-structured interviews with 20 participants: 10 teenagers and 10 parents of teenagers. Our study was approved by the Carnegie Mellon University Institutional Review Board.

3.1 Recruitment and Confidentiality

We recruited participants in and around Pittsburgh, PA by advertising a study on "privacy attitudes" at high school extracurricular activities, through word of mouth, by posting flyers, and on Craigslist. We recruited only teenagers currently attending high school (9th through 12th grade) and parents or guardians of teenagers within that range. To avoid potential biases of interviewing teens and parents drawn from different populations, we required that a teenager and a parent from each household both volunteer to participate in the study. In round-robin fashion, we then selected either the parent or the teen from each household to participate. Although interviewing a teen and parent from the same household would have been interesting, we felt that allowing other family members to know the precise topics discussed could lead to embarrassment or harm after the interview for participants, particularly teens.

Beyond interviewing only one member of a household, we took additional precautions to protect teen participants' privacy. Our recruitment documents and consent form were intentionally vague, noting only that the interview would cover "whether or not teenagers have a right to privacy" and what such a right would entail. We avoided choosing quotes for this paper that we felt would identify particular participants. Furthermore, parents accompanying teens to the study were required to leave the interview room after completing the consent form. The audio recordings and transcriptions were password-protected and not accessible to anyone other than the researchers and transcribers.

We conducted interviews from November 2013 to March 2014. For their participation in our one-hour interview, we compensated participants $30 in Amazon.com credit.

3.2 Interview Procedure and Structure

Interviews were led by one researcher while at least one other researcher took notes and asked follow-up questions. The structure and topics of our interview scripts for parents mirrored those for teens. We began each interview by obtaining consent and explaining the study's purpose.

The topics of the interview included household demographics, technology practices in the household, and the decision-making process regarding the use of new technologies. We also asked teens about their digital personal space. To contextualize a participant's discussion about technology privacy, we also asked about each household's practices regarding physical privacy (e.g., the privacy of a teen's bedroom) and social privacy (e.g., parental notification when a teen goes out with friends). We concluded the interview by asking about teens' general privacy rights. Throughout the interview, we asked follow-up "why" questions for all responses that noted a privacy attitude or privacy decision.

We iteratively adapted our interview script based on previous interviews. The appendix contains our final interview script, which we used for the final eight participants (parents P7–P10 and teens T7–T10). In our initial script, we investigated digital privacy after physical privacy and did not explicitly ask about new technologies. We restructured the interview to emphasize our interest in digital privacy practices. We also originally asked questions about privacy laws, but participants' answers provided minimal insight into the research questions enumerated in Section 1.

3.3 Analysis

The researchers met multiple times during and after the interview process to review their notes and recollections of the interviews and to identify potential themes that warranted investigation in a more structured way during the coding process. These meetings also led to iterative updates to the questions asked in the interview in order to more fully investigate topics discussed by our earlier participants. After the final interview, the researchers met and collaboratively developed a draft codebook containing 88 codes within 15 categories based on their notes from the interviews and previous review meetings. For example, the categories of codes included reasons why teens have a right to privacy, areas of a teen's possessions that are considered off limits, techniques parents use to monitor their teens, and analogies used to compare the physical and digital worlds.

We transcribed each interview to facilitate coding and analysis. A research assistant used the draft codebook to code all of the interviews. We instructed the coder to modify or add codes as necessary to capture anything participants mentioned that was potentially relevant to understanding privacy attitudes or decision making. Following this first round of coding, the researchers and coder met to discuss the coded interviews. The coder had added one code, while eight codes and one category were never used. Realizing that some of the codes were ambiguous in practice, we added 27 additional codes in 6 additional categories. We deleted one

category and ten codes, eight of which were never used.

Using this revised codebook of 106 codes in 20 categories, the coder went back through each interview and revised the codes. A second research assistant independently coded the interviews using the same codebook. The coders had 54% agreement (Cohen's $\kappa = 0.53$). The relatively low agreement appears to result from the large number of codes, some of which the coders felt overlapped conceptually. The two coders met to discuss discrepancies and reached consensus on all codes. We use these consensus codes in all analyses.

3.4 Limitations

Our participants are not a representative sample of the residents of Pittsburgh or any other population. However, our participants did come from a variety of cultural and socioeconomic backgrounds. Participants' families included teens in public, private, online, and homeschool situations.

Since we required parental consent for all teenage participants, we may have excluded teenagers whose parents were unwilling or unable to accompany their child to the interview. This restriction may have disproportionately impacted children of single parents or with troubled familial relationships. As our study was intended to obtain qualitative and anecdotal data from participants and not generalize to a larger population, we accepted this bias.

4. RESULTS

After presenting an overview of participant demographics, we contextualize participants' privacy decisions by discussing their ideas about teens' privacy rights. Most parent participants felt that teens deserved privacy, albeit in a limited fashion. Surprisingly, many teens agreed that teens have only a limited right to privacy from their parents.

We then summarize participants' attitudes toward privacy in the physical world. While both parents and teens discussed providing notice before entering a teen's bedroom, most members of both groups felt that a teen's bedroom was not necessarily a private sanctuary. In fact, the benefits of parental laundry delivery appeared to outweigh privacy costs. However, as we detail in the subsequent section, teens considered text messages on their phones to be private, yet many parents felt it ethical for them to look through their children's text messages. In the final section, we unpack parents' decision-making processes, finding that in the understandable absence of technical expertise, parents made privacy decisions for their teens by drawing false analogies to the physical world or outdated concepts.

4.1 Participant demographics

We interviewed ten parents of teenagers (4 male, 6 female) and ten teenagers (4 male, 6 female). Table 1 summarizes participants' demographics. Our teen participants included 3 freshmen, 4 sophomores, 2 juniors, and 1 senior. All children in P8's household are homeschooled, while T3 attends high school online. The teenage residents of all other households attend traditional schools, including a mix of public, magnet, and parochial schools. We instructed parents to base their responses on their high-school-age children.

Two parent participants (P2, P3) and two teen participants (T2, T4) live in single-parent households due to divorces, while P4 is a widow. Both P5 and P6 live in two-adult households with partners who are not biological parents of the teenage children. All other participants live in two-parent households where both parents are the teenagers' biological parents. Two households (P8 and T3) had other children who attended college and spent most of their time on campus; Table 1 does not include these non-residents.

	Gender	Age	Grade	Children in household
P1	F	—	—	17/M
P2	M	—	—	14/M, 17/F
P3	F	—	—	4/M, 5/M, 6/M, 14/F
P4	F	—	—	14/F, 16/M
P5	F	—	—	12/M, 15/M
P6	M	—	—	15/M
P7	F	—	—	13/F, 15/M, 16/F
P8	M	—	—	14/M, 16/F, 17/M
P9	F	—	—	13/F, 15/M, 17/M
P10	M	—	—	12/M, 16/M, 18/F
T1	M	14	9	14/M, 16/M
T2	M	14	9	14/M
T3	F	15	10	15/F
T4	F	16	10	4/M, 13/M, 16/F
T5	F	18	12	18/F
T6	F	15	9	15/F, 17/F
T7	M	17	11	17/M
T8	F	16	10	16/F, 18/M
T9	M	16	10	16/M, 17/M
T10	F	16	11	16/F

Table 1: Study participants' demographics. Teens are identified by T and parents by P.

4.2 Teens' right to privacy from their parents

Most participants said that teens have some right to privacy from their parents. However, eight teens and eight parents expressly stated that this right is limited. Furthermore, nine teens and all ten parents indicated that parents would be justified in overriding a teenager's right to privacy in an emergency. For example, P6 stated, "I don't know if that is a right or not...they are not necessarily required to share everything with parents...It's not like in the Constitution."

A few participants expressed more rigid views of teen privacy on both ends of the spectrum. Only one teen said that teens should have complete privacy from parents. P5 was the lone parent who agreed, saying, "Anything that they're doing in private is not really my business if they'd not want it to be, and I'm okay with that." Conversely, P2 acknowledged that his children would be surprised "that I feel I should have complete access" to their lives.

4.2.1 Why teens have a right to privacy

Participants, even those who did not think that teens had an overall right to privacy, volunteered many reasons why teens would have a right to privacy. As shown in Figure 1, common themes were trust and teens' inherent need for privacy. Participants also mentioned the importance of giving teens personal space, giving the teen respect, supporting the teen's comfort, fostering a sense of responsibility and independence, and acknowledging privacy as a human right. While seven parents mentioned reflections on the parent's own teenage years, only one teen mentioned this factor.

Ten parents and nine teens said that teens' right to privacy derives from parent-teen trust. In a representative response,

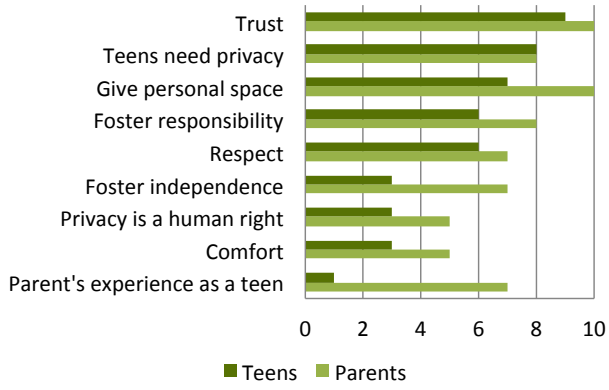


Figure 1: The number of parents and teens interviewed who mentioned each reason why teens should have privacy from their parents.

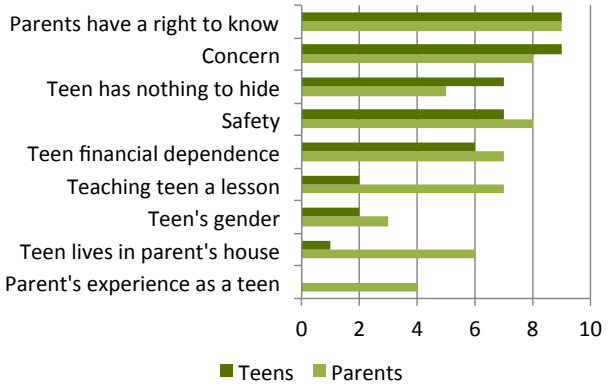


Figure 2: The number of parents and teens interviewed who mentioned each reason why teens should have limits to their privacy from their parents.

T3 said his parents respect his privacy because "they trust me a lot." P4 discussed the importance of earning trust as a prerequisite to earning privacy, explaining, "It's a matter of making me believe in you, making me trust you."

Five parents and three teens said privacy was a human right deriving from dignity. Parent P5 ardently expressed this belief, saying of her sons, "Teenagers are people and everybody has the right to privacy. And just because I gave birth to them and parent them and am responsible for them, doesn't mean that I get to control everything about their lives. Part of teenager-hood is going apart and finding your boundaries, and if I don't let him have any boundaries separate from me, then it's going to make it a lot harder to find his own person...It will affect his life in detrimental ways."

Six teens and seven parents suggested granting privacy was a sign of respect. P3 tied respect to her own experience when she said, "I believe that we should give our kids certain signs of basic respect as is age-appropriate. So if I see [my daughter] is healthy and well-functioning I don't see a need to just go into her room arbitrarily. Just like when I was a teenager I didn't particularly like that."

4.2.2 Why teens do not have a right to privacy

Participants also noted reasons why teens should not have privacy, as shown in Figure 2. Common reasons were a parent's "right to know" and parents' concerns, particularly safety concerns. Participants also said that teens have nothing to hide, teens who depend financially on a parent are obligated to share information, and that teens of a particular gender are more vulnerable and thus do not have a right to privacy. Six parents mentioned that teens in "my house" do not have a right to privacy, four parents mentioned that a parent's own transgressions as a teenager compelled them to look into what their teens were doing, and seven parents mentioned that taking away privacy rights was important when they needed to teach their teen a lesson. Notably, no more than two teens mentioned any of those three reasons.

All but one parent and one teen concurred that parents had a right to know things about their teens because of parental responsibility. P2 felt it ethical to view his children's devices and accounts based on his responsibility for their welfare. He explained, "You're responsible as a parent for them...You need to be aware until they turn 18." P8 stated more generally, "Teens do not have a right to privacy [because] parents are still responsible for their children."

Seven teens, along with eight parents, expressed the need for parents to limit teens' privacy when safety was at stake. P6 displayed reluctance to look through his son's messages unless he was concerned about his son's safety, saying, "I would need to really feel like that violation of his privacy was outweighed, you know, that his safety was more important."

The question of who was paying the bills was also important in determining privacy rights. Six teens and seven parents indicated that teens' financial dependence on their parents minimizes their right to privacy. T7, T9, and T10 all echoed that when parents pay for a teen's education, they have a right to know how the teen is performing academically. P1 expressed dismay that FERPA would bar access to her son's grades once he turns 18 even if she pays for his education. She said, "I think that's wrong...If the parent's paying for it, I want to know what's going on."

Financial dependence also drove the sentiment expressed by six parents that, because teens live in houses owned by their parents, they have fewer privacy rights. T8 said, "It's their house, so they can do what they want." Similarly, P2 expressed his right to enter any room in his own home: "It's my house...If I need to go in there, I'm gonna go in."

4.2.3 The boundaries of privacy rights

Nearly all parents noted boundaries to teenagers' privacy rights, often explaining that these boundaries are fluid. P6 wrestled with these limitations: "I will do my very best to honor [my son's] privacy, but if at the end of the day I need to do something that violates [his] privacy because I feel like it's the right thing to do...then I will violate the shit out of his privacy...That's my responsibility as a parent."

P4 explained the difficult balance between privacy and control: "At one point [my son] called me controlling. I don't think I'm controlling. I think I'm protective." She expressed her struggle by saying, "I wanted to...not be controlling, but I still wanted to have some control."

Some parents noted areas as expressly permissible for parents to access. Both parents and teens commonly noted grades in school as non-private. P7 was dismissive of teens

keeping grades private: "Grades? No, that's not privacy to me." Many teens noted that their schools automatically shared grades with parents and that they had neither the ability nor the right to keep grades private from parents.

Parents also noted situations they would consider violations of their children's privacy. P2 said, "If my daughter likes some boy at school and she doesn't want to share that with me, that's fine." Similarly, P5 described why looking through her sons' digital files was inappropriate: "I don't think there's any reason to. I mean, if the teenager agrees to it, but only then. And that's often questionable because I think it can be very easy to coerce them into agreeing."

While many parents reserved the right to override teens' privacy rights, our interviews suggest they do so infrequently in practice. Asked whether she has the right to read her daughter's correspondence with friends, P3 said, "I do, but just because you have the right to do something doesn't mean that it's morally the right thing to do."

4.2.4 Age, maturity, finances, and college

Participants noted that privacy rights are not static. They commonly felt older teens should have different boundaries and privacy expectations. Privacy rights evolved based on age, maturity, financial independence, and starting college.

Privacy rights increased with age according to seven teens and all ten parents. P2 described how his practices changed as his children grew older: "I, as time went on...allowed them to make their own choices." Six parents and six teens also mentioned maturity. P7 explicitly distinguished maturity from age, saying, "It really depends on the maturity of the kids. And not necessarily the age."

Participants had nuanced views around privacy changes when teens turned 18, the legal start of adulthood in the United States. T5 acknowledged the legal boundary at age 18, saying, "I think when you turn 18, your parents even owe it to you in a way to give you more responsibility." P4 also acknowledged this boundary, saying, "He's going to be eighteen, so I don't really have any say at that point." However, she also lamented, "I think eighteen is young." Surprisingly, few parents or teens expected that teens should obtain full privacy rights on their eighteenth birthday.

Six teens and eight parents cited increasing financial independence as a factor impacting privacy rights. P4 indicated that she would give her son more privacy when he started financing his own phone: "At some point he's going to pay for his own phone and stuff, and...there should be trust there so I shouldn't have to look at it."

4.3 Privacy in the Physical World

Parents were generally willing to carve out private space in the physical world for their children. Privacy in the physical world did have its limits, though. P1 directly addressed the superficial tension between teens and parents regarding rules: "There's some resistance, but I know in the end [my son] appreciates me and loves me for it."

We found that parents generally let teenagers keep the door to their bedroom closed, except when significant others were visiting. All parents felt entitled to enter their children's rooms when their children were not there. As long as parents were not snooping, most teens agreed. Teens appeared to consider few physical areas private. Most parents had rules and restrictions about their children's social lives. All teens were required to notify their parents of their physical location at all times. While these requirements did cause some parent-teen tension, both parents and teens generally agreed that such practices were reasonable.

4.3.1 Bedrooms

We found that parents generally treated teens' bedrooms as somewhat private, giving the teens personal space, yet did not feel like they should be restricted from entering. While teens did not approve of the relatively rare practice of parents snooping around their room, they felt that the benefits of having their laundry or other tasks done for them were valid reasons for their parents to enter their room.

We found wide acceptance of the practice of teens keeping their bedroom doors closed for privacy. P2 said he permitted his children to keep their bedroom doors closed because "that's their space." A few parents gave their children privacy in their bedrooms, yet explicitly noted that they would still go in if they wanted to. As P3 explained, "Mom reserves the right to check on any of her children at any time."

All parents and all but two teens indicated that parents knocked or otherwise notified their children before entering their room. Respect often drove this decision; P1 explained, "It's his private [space], it's his domain. Well, not domain, but just out of respect. I'd expect the same." Generally, parents and teens used knocking or other advanced warning to avoid awkward situations. For instance, P5 explained, "Since the door is closed, there are potential things I could be walking in on that neither of us want to know about."

All parents felt comfortable entering their children's bedrooms when their children were not there, and all teens except T4 said their parents enter their room when they are not there. None seemed particularly troubled by this practice as long as their parents had a reason. For instance, T7 felt it was acceptable for his parents to come in "to get my laundry. That's pretty much it. Or make my bed."

A handful of parents interviewed did think it appropriate to snoop through children's rooms. P1 noted that she enters her son's bedroom multiple times a week "to snoop. It's my house and I'm gonna go in that room whenever I want to." Despite this snooping, she did not feel like she was violating her son's privacy. She explained, "Hell, there could be a mad man living in the room, how would I know? I could see Dr. Phil, 'Well, you never went in your son's room, huh, would you now?' Ya, I respect his privacy, yes I do."

Although most parents and teens generally considered unprovoked snooping a privacy violation, only a few participants felt particular areas of the physical bedroom should be off limits to parents. Some parents who mentioned specific locations noted that these policies were hypothetical. For instance, P1 said "if [her son] had a diary, I wouldn't look through that." No participant other than P3's daughter actually kept a diary.

Instead, both parents and teens most commonly mentioned cell phones and computers as off limits to parents. Seven parents each mentioned cell phones and computers. While nine teens mentioned cell phones, only five mentioned computers. Figure 3 enumerates the locations and devices participants suggested were off limits for a parent.

4.3.2 Privacy in teens' social lives

All participants except two parents and three teens noted that the teenagers in their household had restrictions on their social lives, most commonly curfews or restrictions on

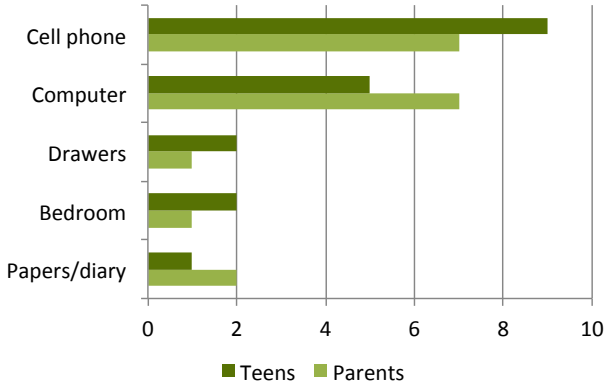


Figure 3: The number of parents and teens who said different areas were inappropriate for a parent to go.

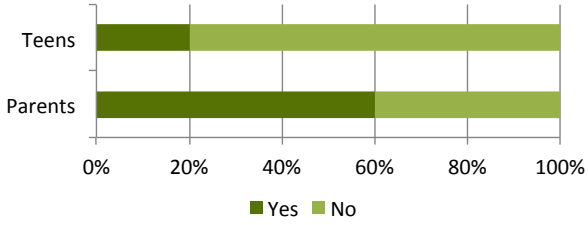


Figure 4: The percentage of parents and teens who felt it ethical for parents to look through teens' text messages. Teens strongly opposed this practice.

overnight visits. Whereas T10 was representative in saying "[My mom would] never let me sleep over at a guy's house, or let a guy sleep over at my house," T4 was similarly representative of the flexibility most teens created for themselves, saying, "Oh yeah, [I violate my parents' restrictions] a lot. I'm always late on curfew."

All participants, even those without restrictions, noted that the teens in their household needed to notify their parents in advance about where they were going. There were some complaints about required notification, but they were limited. P1 discussed her son's objections, saying, "He says I'm always calling him...And he wants his own personal space." Surprisingly, all of the teens we interviewed felt the notification process was reasonable, though annoying.

We also investigated attitudes about teen dating and romance, particularly concerning privacy. While participants had a range of views on the appropriateness of teen dating, these discussions provided little insight into privacy decision making. Most commonly, parents wished to be oblivious to their children's sex lives. As P4 said, "It makes me a little bit 'ew'...I'm not sure that I want to know." P1 explained, "I don't even want to think about it...That's disgusting."

4.4 Teen Privacy in the Digital World

We investigated teens' attitudes towards digital devices, including laptops and phones. Teens largely expressed that their digital spaces were personal and private. We also note the prevalence of teens using laptops solely for schoolwork, rather than recreational browsing, as well as the prevalence of teens using texting as a primary communication channel.

4.4.1 Devices

Teens felt strongly that phones were private devices that parents should not access. In contrast, laptops were less private and primarily for schoolwork. T9 said, "People don't have the right to go through [my phone]." Similarly, T2 expressed annoyance at his parents "constantly searching my phone," which he thought demonstrated a lack of trust. P9 was aware of the value her children placed in phones, stating, "no one should be in each other's phone...invading other people's rights." Few teens used computers for socializing. T8 spoke for many of our participants when she said that phones were "more private" than computers.

Many participants relegated laptops and personal computers to schoolwork only. As T5 explained, "A lot of schools will provide you with a Chromebook or something of that sort." This led some teens to distrust the school-provided laptops because, as T4 put it, "I think that they can go in and check [my activities]." As a result, many teens chose not to use their computers for anything other than schoolwork. T7 said, "I don't really use my laptop that much, just for like schoolwork and stuff like that." T8 concurred that her computer was used for "mostly school," while T10 stated that "the computer just has school documents on it."

Most parents also observed that computers were primarily for schoolwork. Of her son's laptop, P2 explained, "He hardly ever uses it, except for schoolwork." Monitoring and content-control software on school laptops seemed to be a significant reason for teens' minimal use of laptops. As P5 describes: "my older one has a school provided laptop that is very locked down, and he really only uses it for school."

4.4.2 Texting

Many teens used text messaging for private communication. Our teen participants repeatedly echoed this thought, with T2 saying, "[Parents] looking at my texts...feels like an invasion of privacy," and T10 explaining, "Texts are more private because that's where I talk to my friends."

Some parents had also observed that teens relied on texting for private conversation. P9 observed that her children "use their phones for socializing and I don't feel the need to get involved in that. And they don't want me to." However, many participants said that parents in their household had no qualms looking through text messages, such as when T2's mother punished her son by looking through his texts. He explained, "I lost her trust and she decided to look through all my text messages on my phone." In some households, parents monitored teens' texts even more regularly. P10 and his wife routinely checked their children's phones for "texting, anything they can access with that," and he expressed that he would prefer to be "checking [texts] more consistently." While many parents felt it acceptable to monitor texts, very few teens agreed, as illustrated in Figure 4.

4.4.3 Social Media

In general, teens' reported interest in traditional social media seemed to be waning compared to past studies. We asked specifically about teens' use of Facebook. Signing up for an account often came "with the provision that if you have a Facebook account, you friend your parents," as explained by P8. Perhaps in reaction to this, teens had moved

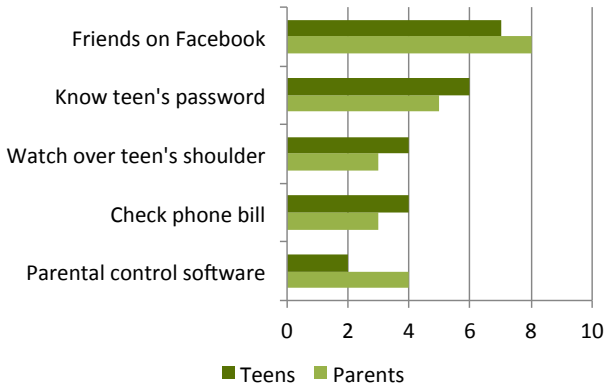


Figure 5: The number of parents and teens we interviewed who reported the adults in their household employing different types of technological monitoring.

away from Facebook for personal activity and correspondence. Besides running a few group pages and talking to family, T3 admits, "I don't really do anything on Facebook besides just like, checking every once in awhile." P6 observed, "I think that Facebook is kind of trending away with the younger generation, especially when his parents and all his parents' friends are on there."

While few teen participants said they regularly used Facebook, many of them did report using Instagram, Snapchat, Vine, and Twitter. However, teens' shift from Facebook to these new services was not yet on most parents' radars. None of our parent participants were familiar with Instagram or Vine, while only one (P6) had ever heard of Snapchat. P6 was not sure, however, whether his son used Snapchat.

4.5 Decision making

In this section, we first examine parents' decisions about monitoring and restricting their teens' use of technology as examples of privacy decision making. We then unpack parents' decision making about their teens' privacy. We discuss how parents rely on their own experiences as a teenager to make parenting decisions about technology, even though participants thought today's world differs substantially from the world of 25 years ago. We also evaluate the extent to which participants conceived of privacy similarly online and in the physical world, which was a source of parents' misunderstandings of how private teens consider their devices.

4.5.1 Decisions about monitoring

Parents utilized a variety of methods to monitor their teens' online activities. As shown in Figure 5, parents commonly relied on being Facebook friends with their teens, knowing some of their passwords, or looking at their teens' phone bills. Some parents placed computers in public areas of the house to watch over their shoulders, while others surreptitiously monitored teens' activities. Four parents and two teens reported that parental controls had ever been used in their family, while four parents (P1, P6, P9, P10) simply required their children to show them their devices.

In a few households, computers were kept in a public area so that parents would know what their teens were doing. When describing her 14-year-old daughter's computer use, P3 stated, "Her computer's kind of in a public room where mom can see." T1 described how his parents would observe by listening, saying, "Sometimes they come and just stand there and they don't say anything. Normally, they're just listening to my Skype conversation." A few parents also monitored surreptitiously. For instance, P7 said, "Sometimes it happens that I check [my kids' browsing] history."

Four parents and two teens mentioned parental control software. P8 used parental controls primarily to restrict certain content, but had dabbled in blocking social media: "Basically it was a filter for, just, stuff we thought might be offensive sites...earlier on we kinda limited some social networking, because we wanted to get a sense of who they were communicating with." Notably, however, most of the families that had used parental controls had since abandoned them due to the frequent false positives. In other cases, the families had no idea whether the parental controls were still active. A few participants reported using parental controls on mobile or portable devices. T9 had struggled against restrictions on his iPod, recalling, "[My Mom] put a password on it to sort of change the restrictions [to prevent downloading explicit songs], so I tried to guess it; successfully, eventually, but by then it didn't really matter."

Parents often paid for teens' cell phone plans, so some had access to records of whom the teen had called and texted through the monthly phone bill. Many parents took advantage of this and reviewed the logs (P5, P6, P7). As T3 reported, "They check which numbers I text." For the most part, teens assumed that parents were primarily monitoring not who the teen contacted, but that teens were not staying up too late (T3 and T9). In T2's case, his mother, who lived separately, looked at his phone bill to ensure that he was not ignoring her. He said, "[If] I haven't talked to her in a couple of days...she checks the phone bill to see if I've been texting people and calling, to see if I'm ignoring her."

Some parents looked through teens' text messages. P5 did not do so, which she felt made her an outlier. She explained, "I have many many friends who have, say, teenagers or pre-teen kids who do think it's absolutely acceptable to take their phone and look through their kids' messages, or limit things, or just read over emails." T4's father would frequently look through her text messages. She developed a strategy to avoid her father's prying, saying, "I leave [my phone] in my room and I'll just tell my dad I forgot it."

A common practice was for teens to be friends with their parents on social media, often as a condition of using the site. As T6 put it, "I'm friends with my parents on Facebook. That's, like, a big thing." P10 dryly remarked of his children, "Yes, [my wife's] on Facebook, much to their chagrin." However, not all teens friended their parents under duress. P6 described his son's Tumblr use, saying, "My girlfriend is a follower of his on Tumblr. And he follows her. It's very out in the open. We're not sneaking up on him."

Five parents and six teens reported parents having access to teens' passwords. Motivations for this practice differed. Sometimes, passwords were necessary to maintain the computers. As P8 explained, "When I need to go in and manage their side specifically, I log on with their password." Monitoring content was a prominent goal for other parents. P10 and his wife regularly checked his teens' computers, using their passwords for "checking for online searches, checking email, Facebook, social networking, that sort of thing."

Many parents reported having access to passwords, yet not using them. T10's family keeps a written list of passwords that all members of the household can access. Some parents helped teens set up computers or accounts and had their teens' passwords as a byproduct of that process. For instance, P5 said, "I did help them set up their Gmail accounts years and years ago. They probably haven't changed their passwords, so I probably still have them. But I haven't logged in in five years or whatever." P3 explained having her daughter's password was a safety measure, saying, "I just have the password. To me it's kind of a safety thing. I have jumper cables in my car...I hope not to have to use them tonight, but they're in there just in case."

Deleting information, such as browser history or text messages, was a common tactic among teens to avoid exposing private content to parents. Surprisingly, a number of parents expressed that they wanted their teens to delete things. For example, P4 lamented, "I'd think to myself, why didn't you delete it?" Teens expressed that they tried to be clever about covering their tracks. T4 explained, "I try to delete some of [my text messages] so it's not really obvious." Managing and routinely clearing questionable data took a toll on some participants. As T9 admitted, "I've watched pornography...At the time that I did, I was really a lot more paranoid about search history and stuff."

4.5.2 Decisions about restrictions

When parents attempted to regulate teens' technology use, they turned to non-technical methods. Parents sometimes took devices away, imposed time limits, or specified where devices could be used. As punishment, parents took away devices and shut off Internet access. When P9 wants to discipline her children, she "will take the phones away when I feel they're acting disrespectful." As T6 admits, "My dad turned the Wi-Fi off my house at one point." Parents usually imposed time limits verbally. T10 explained, "We'd play games and they'd say, 'Okay, only 15 minutes.'" Other families required devices to be used or not used in certain areas of the house. T6 explained, "They don't usually let us have laptops in our rooms."

4.5.3 Parents' own teenage years

In determining what policies to set for their teens' privacy, parents commonly used their own experiences as teenagers. For instance, P1 explained, "I try to think back when I was his age." She actively gave her son some private space, even though she decided that snooping in his room was not a violation of his privacy. On the other hand, P5 explained that she emphasized being open with her children, lamenting that her mother "never started the conversations."

Other parents mentioned their own transgressions as informing their parenting decisions. For instance, P6 said, "When I was fifteen, I totally would have broken all those restrictions." Similarly, two other parents mentioned their experiences as teenagers hiding marijuana from their own parents. Amusingly, P4 lamented her own children's inability to hide their tracks, saying, "These kids today. When I snuck out of the house at his age, I made sure that I came in and left and didn't get caught!"

4.5.4 Differences today

While parents' own experiences are crucial to their decision making, all participants noted many ways in which being a teenager today is different than it was 25 years ago. In addition, except for P8, all parents said their view of teen privacy is different from how their parents viewed it when they were teenagers themselves. Technology played a major role in these changes, and the use of technology was starkly different. For instance, P6 mentioned that his son started using a computer at age 2. In contrast, he said, "When I was fifteen, we had an Apple IIe computer at home...I honestly couldn't do anything with it."

The most salient difference was a tension relating to teens' freedom. While modern teens have the freedom to access huge amounts of information, they lack the freedom to disappear from their parents (P1, P5, P7, P9). The expectation is that they are always connected. As P4 explained, "I have to know where he's at. If I call him he has to answer." Similarly, P5's kids had "just gotten smart phones and one of the agreements for that was that I need to be able to get in touch with them whenever I need to." Some teens were cognizant of the implications of cell phones. As T5 said, "There comes a lot with a cellphone, in the sense that you can be reached at any time. Or be bothered."

Parents contrasted modern expectations of constant availability with their own childhoods. P7 reminisced, saying, "My parents didn't know where I was for hours...I couldn't call." Similarly, P9 recalled, "I'd say I'm going to New York...[and] come back like eight hours later. Did they ask where I was, what I did? No! I was back."

For many parents, technology thus became a means of control (P4, P5, P6, P7, P9). P7 explained, "The reason why we both have the phones is because we parents want to actually control them. So, at least 50% of the reason was for us, not for them." P6 focused on the use of technology in schools to keep tabs on his son. He said, "Thank God for technology...I do look at his grades and his missing assignments...It's kind of like that whole panopticon thing."

Some parents felt technology has made teens' lives much more complex (P4, P9) and dangerous (P2, P7, P10). P10 said, "There weren't as many issues as there are today for teens for a privacy issue to arise...The biggest problems in schools were chewing gum, and these days it's weapons and drugs and rock & roll, alcohol." Teens tended to be somewhat dismissive of this viewpoint. T8 explained, "When my parents were my age, they tell me about how they walked everywhere and how they could keep their doors unlocked and we can't do that today." When asked why this was no longer the case, she snarkily replied, "Because I could get abducted or something, I don't know." T9 instead characterized the generation gap as one of access. He said teens now "do things much more efficiently, like setting up a party...It's kind of like what they had, but for us it's on steroids."

Some parents (P2, P5, P6, P7, P9) expressed shock at the extent of teens' lives that occurs online. As P6 explained, "[My son] spends a significant portion of his life online. He really does. And I think most kids do." He contrasted this state of affairs with his own childhood: "The things that were private when I was fifteen were my bedroom and what was going on in my life." Teens also recognized parents' confusion at the way teens communicate. For example, T8 said, "They think it's weird that I'm on the computer a lot, but it's just something that this generation does."

One parent noted that people her age are the first to have experienced stark generation gaps between parents and teens. P9 explained that, for her own parents, "the big

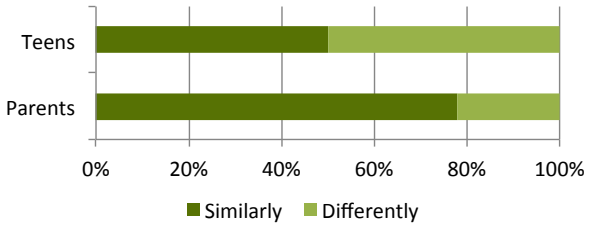


Figure 6: The percentage of parents and teens who said they think similarly or differently about privacy in the physical world and privacy online. We exclude one parent and two teens with whom we did not discuss this topic.

thing [for my parents]...was like smoking and drinking and playing cards...I think that their lives were so similar to their parents' lives, like there wasn't a big culture break...I was educated in a lot of the liberal views and sexual freedom...Honestly, I was a lot wilder than my kids ever will be." She drew another distinction with her children's generation. She said, "I worked from age 12 because life was boring...[Now] I feel like we can do so much with a phone: we can look up, we can research, we can read books, we can talk to people." As a result, she had to kick her "oldest son's butt to work and he's like 'Why? I have a phone. I have access to a car. I have friends. I get good grades. Why would I need to work?'"

4.5.5 Online vs. physical

A large part of the gap between parents' and teens' privacy decision making process appeared to be predicated on whether they thought similarly about privacy online and in the physical world. Excluding three participants who did not discuss this topic, 7 of 9 parents said they thought similarly about privacy online and in the physical world, whereas only 4 of 8 teens said the same (Figure 6). It is less surprising, then, that teens and parents differ in opinion about privacy for digital devices.

Many parents felt that the human characteristics underlying the physical world applied equally to the digital world. P3 explained, "I didn't have a cell phone. I didn't have Internet back then. It was kind of like a different world. But as far as respect goes, basic respect is always gonna be basic respect." Similarly, P6 offered, "Just because they're not sitting in front of you doesn't give you the right to talk in a way that you wouldn't talk to someone who was right there in front of you." Other parents had not given potential contrasts between the digital and physical worlds much thought. When asked detailed questions about parenting practices, P2 was surprised to conclude, "It seems that I do give them more space online than I do physical."

P6 drew a direct parallel between his son's privacy in the real world and online. Early in the interview, he mentioned his son had a small chest in his room. P6 said his son "had a lock on it for a little while, not actually locked, just kind of hanging on the thing. It was an interesting symbolic demonstration of 'this is off limits,' even though it clearly wasn't, because anybody could just take the lock off." P6 chose to treat the chest as his son's private space and not snoop. When later discussing how he knew his son's computer passwords, P6 explained, "I want the password in the same way that I [would] want a key to the lock on his footlocker...I'm not going to snoop around in his stuff...[But] if I had cause to think I needed to look at that stuff, I want to be able to do so at my convenience."

P7 also drew parallels between her son's behavior on the computer and her experiences as a teenager in the physical world. When her son deleted his computer's browsing history, P7 found it "really smart but also suspicious. I mean, you clean the history because you want to hide something. But then I had a second thought and I said, well, I tried to actually create a similar situation when I was a teenager. I remember that I wrote down notes, not really a diary, but some kind of personal notes. And I could have hated my parents after reading that stuff. So I try to respect this kind of private life for whatever it is."

Interestingly, the parents and teens who distinguished between privacy online and privacy in the physical world had diametrically opposed views about the relative danger of these contexts. Parents felt that the physical world was filled with friendly faces, yet online was filled with strangers. P5 explained, "My concern for privacy online is much more of protection from other people...Like if they close their bedroom door, the only risk to me looking inside is that I'm looking inside. Whereas if they have a Facebook account or whatever, and they don't close the door properly, then a billion people can look inside." P7 similarly asserted that online, "your stuff is available or reachable by a much bigger context. So if you publish something, it's not just your circle of friends or family, it can really go to the world. So the impact is ten times, one hundred times bigger."

In contrast to the parents, the teens we interviewed felt their online world comprised their friends, whereas the physical world was inhabited by strangers. As T8 explained, in the "physical world, the majority of the people I see are strangers, so I don't really worry about them thinking about what I'm doing. But online, like the people that follow me, I know them personally. So I think what I do will kind of affect them more in how they see me."

The teens we interviewed also had trouble understanding the online dangers their parents emphasized. In fact, the teens felt that there were fewer possible consequences online than in the physical world. T7 explained, "If someone finds out that you did something in life, you can get in trouble at school or get in trouble with your parents or something like that. But online, there's not that much stuff that you can get into trouble." With one exception—T5, who noted concerns about online hackers and cyberbullying—the teens felt that their school assemblies about online safety were excessively alarmist. T8 characterized the latest such assembly as "kind of boring, so I kinda like just zoned out." As a result, teens felt that parents misunderstood the decisions they were making about online safety.

The teens did note some exceptions to the general safety they perceived online. T6 noted that, in the absence of actively deleting information, "online stays there forever." Furthermore, teens made a specific exception for Facebook, which they felt was the one area where their online life intersected their life in the real world. T10 said, "It makes sense to me why [my mom would] want to monitor my Facebook because people talk about how when you're looking for a job they'll look on your Facebook."

4.5.6 Misunderstanding teens' private spaces

One major disadvantage parents had in their decision making process was an incomplete understanding of the technologies their children were using daily. Half the parents we interviewed said they struggled with technology (P2, P4, P5, P7, P9). P4 told us she "just realized that you're able to go online" with gaming consoles. Describing Android unlock patterns P6 said, "I watch him do it sometimes and I still don't understand." P2, who works in the tech industry, explained that the mother of his children was reluctant to let the kids have email accounts because she is "more paranoid about things she doesn't understand." Yet he also struggled when trying to comprehend his son's "36 virtual friends."

Unsurprisingly, teens then felt their parents failed to understand modern communication. T6 lamented, "They think that you're behind a screen, so you're cutting yourself off from the world. But I don't think that. I think you're talking to people." This tension clearly manifested itself regarding text messages. Whereas T5 happily noted texts as the default communication channel, P1 complained that texting "makes me mad. I want to hear [my son's] voice."

Beyond teens' reliance on text messaging, the parents we interviewed struggled to understand the private nature of those conversations. For instance, even though T6 often deletes text messages, she said, "[My parents] want me to think before everything I write, even in a [text] message." The impermanence that teens attributed to phone conversations carried over to apps. Even though she was aware of the ability to take screenshots, T6 considered Snapchat to be her most private method of communication. Only one parent (P6) had heard of Snapchat; none had ever used it.

Whereas teens relied on the ability to delete messages on the phone, parents felt that digital communications were uncontrollable once sent. P10 said text messages "can be forwarded. They can be copied. Other people can find out about them." He felt that the only type of private communication between friends was "a written note [or] getting alone with them [in a] room...the old fashioned ways." Similarly, P3 believed teens communicate privately via notes, as was the case during her childhood. She said, "Teenage girls, I was one of 'em once, pass notes to her friends in her school." As a result, she did not consider her daughter's phone to be private and had configured her daughter's phone to communicate only with whitelisted numbers.

Some of parents' unfamiliarity benefited teens. T8 explained, "My dad, he's bad with technology. But my mom could adapt if she wanted to. I don't think she really cares to...She mostly just pays attention to Facebook," which was convenient for teens since none of our teen participants considered Facebook to be particularly private. In contrast, T8 thought it "would be weird" if her mom wanted to follow her on Twitter. Similarly, T10 happily mentioned, "I don't even know if my mom knows what Snapchat or Instagram is."

A major difference we observed between parent and teen participants was their understanding of what types of private spaces were most essential. Most parents thought allowing their teenagers to be alone in their bedrooms with the door closed was sufficient private space. Putting herself in her son's shoes, P1 said, "This [bed]room is my world. I can listen to my music, go on the computer, do what I want." P5 noted that "everybody needs a space that they can go to that they can just be private...Since the house is mine, the bedroom is really the only space [teenagers] have." By a similar thought process, P6 noted carefully avoiding looking "under the mattress" when he needed to search for bedbugs in his son's room because "when I was a child, that was a place where you hid things away from your parents."

However, teens generally did not hide things under their bed; they hid them in their phone. Many of the parents we interviewed did not grasp the importance of cell phone privacy for teens. Even P6, who generally felt "it's not ethical to go through anybody's text messages [because]...it's the equivalent of digging through somebody's drawers," struggled with teens' phone privacy. He later noted, "if the day comes that I really want to look at his phone, I could."

4.5.7 Parents' struggles evaluating privacy

In the end, all but one parent said they struggled making privacy decisions for unfamiliar technologies. When asked how he decides what rules to adopt for new technologies his son is using, P6 said, "I kind of make it up as I go along." He further explained that his son "has access via the Internet to things, materials—explicit materials in particular—that when I was fifteen, you just didn't have access to...And that does pose a problem in terms of what does that really mean? But that's an answer that I don't have."

The lack of context caused particular difficulty. P8 simply noted that "the playing field is different." This different playing field left parents unable to evaluate risk; P7 complained, "How can you compare? Like my kids can actually be in the dining room and chatting with somebody in China...The reality is they could actually be in more danger."

The rapid pace of change was an additional confound. P10 explained, "You're comfortable with what you're familiar with. And today things are changing so much that it's hard to get familiar and comfortable with something because there's a new advancement something's new and improved, or there's a whole new way of communicating." Similarly, P4 lamented, "It's overwhelming for me...It's so different from when I grew up...I don't know if I'm too strict or too loose."

5. DISCUSSION

Our interviews with ten teens and ten parents delved into how parents understood and navigated teens' privacy in an unfamiliar world, as well as how teens perceived their parents' decision making. Our findings unveiled a notable disparity between teens' and parents' views of technology, cutting across family dynamics and socioeconomic classes.

In some areas, we found accord between parents and teens. Both groups generally acknowledged that teens had some right to privacy from their parents and that this right was limited. However, as we look toward real-world examples of privacy rights, parents and teens begin to diverge. Many teens felt that their smartphones, containing text messages and apps, were their most personal form of communication. Even when the parents we interviewed expressed a desire to give their teens personal space to socialize with their friends, they anticipated the teens would have an in-person conversation, not use text messaging. As a result, these parents often adopted policies regarding use of technology that clashed with their abstract goals of giving teens private space.

Despite their conflicting perceptions of technology use, both parents and teens were operating in good faith. Communication problems were the heart of the issue. Parents struggled with how to make decisions about technology use—they weren't intimately familiar with many of the technolo-

gies and made incorrect assumptions. Meanwhile, teens were more familiar with the technologies, but were not always able to make responsible and mature choices. In one example, parents frequently required that their teens friend their parents on Facebook as a condition of signing up for the site. While parents felt this was a good way to keep tabs on their children's digital activities, it seems to have caused the teens we interviewed to stop using Facebook regularly. Instead, teens overwhelmingly preferred texting, Instagram, or newer apps for socializing with friends.

The communication gap arising from generational differences and differing perspectives on the role of digital devices and the Internet is a substantial obstacle to parents' decision making. We intend this paper to inform the conversation about how to help parents make privacy decisions for their teens in this technology-filled world that differs starkly from their own childhood. While we did not test specific approaches, our results provide insight into the needs of parents and teens that can help guide developers.

Even though many of the parents we interviewed described struggling with making decisions about privacy for their teens, few of them regularly used parental controls or other digital parenting software. Even the families that had used these tools in the past reported that the trouble of using them often outweighed the benefits. One reason for this non-adoption might be that the tools do not support parents' goals sufficiently. Existing digital parenting software most commonly blocks access to resources deemed inappropriate according to some heuristic. Frequent false positives in this blocking cause frustration and lead parents to disable these parental controls [27]. Other tools are designed to notify parents about their children's activities, such as their location. However, parents sometimes find this approach stifles their children's independence and maturation process, again leading to non-adoption [28].

Our results suggest that there is ample opportunity for tools that inhabit a middle ground between doing nothing and forcibly preventing or conspicuously reporting teens' actions to their parents. Parents who are concerned that they are not doing enough to teach their children to make responsible, privacy-protective decisions when using technology might find value in software tools that encourage, rather than force, certain types of behaviors. This approach of encouraging, or "nudging," users to give more careful consideration to a decision has been applied successfully to a number of domains [26].

Among digital parenting software tools, this approach to software might use heuristics to detect actions that a parent might not approve of and take the opportunity to remind the teenager of the parent's expectations and the teen's responsibilities, yet not block the action. For example, in a field trial of privacy nudges for Facebook, Wang et al. found that visual reminders of a family member being able to view content was effective in encouraging privacy-protective behaviors [29]. The nudging approach to digital parenting software might alleviate parent-teen tensions because teens would still be free to make their own decisions, albeit with guidance and reminders.

Our results can also inform efforts to improve user education around these new technologies. In particular, we observed a major gap in parents' understanding of how their children use new types of devices, apps, and services to communicate with their friends. Unfortunately, much of the discourse in the popular media about these new technologies focuses on worst-case scenarios. Instead, parents might benefit from a better understanding of how the majority of teens actually use apps like Snapchat [19], beyond the fact that a fraction of teenagers use it to send explicit photos. Similarly, the increased understanding of parents' and teens' perspectives that we provide can be used to improve laws like the Children's Online Privacy Protection Act of 1998 (COPPA). While other scholars have noted flaws in the implementation of COPPA [3, 13], our additional perspective can help suggest potential next steps in improving privacy laws.

6. ACKNOWLEDGMENTS

This work was supported in part by NSF grant CNS-1012763 and by a National Defense Science and Engineering Graduate (NDSEG) Fellowship awarded by the DoD and Air Force Office of Scientific Research (32 CFR 168a). The authors would like to thank Richard Shay for his assistance.

7. REFERENCES

[1] K. A. Bamberger and D. K. Mulligan. Privacy on the books and on the ground. *Stanford Law Review*, 63, January 2011.

[2] d. boyd. *It's Complicated: the social lives of networked teens*. Yale University Press, 2014.

[3] d. boyd, E. Hargittai, J. Schultz, and J. Palfrey. Why parents help their children lie to Facebook about age: Unintended consequences of the 'Children's Online Privacy Protection Act'. *First Monday*, 16(11), November 2011.

[4] J. Cheng. What inner city kids know about social media, and why we should listen. Medium, September 25, 2013. `https://medium.com/i-m-h-o/53ea514c9ec0`.

[5] J. T. Child and D. A. Westermann. Let's be Facebook friends: Exploring parental Facebook friend requests from a communication privacy management (CPM) perspective. *Journal of Family Communication*, 13:46–59, 2013.

[6] A. Czeskis, I. Dermendjieva, H. Yapit, A. Borning, B. Friedman, B. Gill, and T. Kohno. Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In *Proc. SOUPS*, 2010.

[7] A. Forte, M. Dickard, R. Magee, and D. E. Agosto. What do teens ask their online social networks? Social search practices among high school students. In *Proc. CSCW*, 2014.

[8] S. Hawk, L. Keijsers, W. Hale, and W. Meeus. Mind your own business! Longitudinal relations between perceived privacy invasion and adolescent-parent conflict. *Journal of Family Psychology*, 23(4):511–520, 2009.

[9] A. Hess. Millennials aren't oversharing on social media. (So what are they hiding?). Slate, October 18, 2013. `http://www.slate.com/blogs/xx_factor/2013/10/18/millennials_on_social_media_young_people_are_incredibly_savvy_about_internet.html`.

[10] S. Livingstone and M. Bober. Regulating the Internet at home: Contrasting the perspectives of children and

parents. In *Digital Generations: Children, young people and new media*, pages 93–113. 2006.

[11] M. Madden, A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton. Teens, social media, and privacy. Pew Internet Report, May 2013. `http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx`.

[12] A. E. Marwick, D. M. Diaz, and J. Palfrey. Youth, privacy, and reputation. Harvard Public Law Working Paper No. 10–29, 2010.

[13] L. A. Matecki. Update: COPPA is ineffective legislation! Next steps for protecting youth privacy rights in the social networking era. *Northwestern Journal of Law & Social Policy*, 5(2):7:1–7:35, 2010.

[14] A. Metzger, C. Ice, and L. Cottrell. But I trust my teen: Parents' attitudes and response to a parental monitoring intervention. *AIDS Research and Treatment*, 2012.

[15] M. K. Nelson. *Parenting Out of Control: Anxious Parents in Uncertain Times*. NYU Press, 2012.

[16] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1), 2004.

[17] S. Petronio. Privacy binds in family interactions: The case of parental privacy invasion. In *The Dark Side of Interpersonal Communication*, pages 241–258. 1994.

[18] J. A. Rode. Digital parenting: designing children's safety. In *Proc. BCS-HCI*, 2009.

[19] F. Roesner, B. T. Gill, and T. Kohno. Sex, lies, or kittens? Investigating the use of Snapchat's self-destructing messages. In *Proc. FC*, 2014.

[20] J. Smetana. "It's 10 O'Clock: Do you know where your children are?" Recent advances in understanding parental monitoring and adolescents' information management. *Child Development Perspectives*, 2(1):19–25, 2008.

[21] J. Smetana, H. F. Crean, and N. Campione-Barr. Adolescents' and parents' changing conceptions of parental authority. *New Directions for Child and Adolescent Development*, 108, 2005.

[22] D. J. Solove. Conceptualizing privacy. *California Law Review*, 90, 2002.

[23] D. J. Solove. "I've got nothing to hide" and other misunderstandings of privacy. *San Diego Law Review*, 44:745–772, 2007.

[24] D. J. Solove and P. M. Schwartz. *Privacy Law Fundamentals*. IAPP, 2013.

[25] B. Teitell. In digital world, kids gain the upper hand. Boston Globe, September 5, 2013. `http://www.bostonglobe.com/lifestyle/style/2013/09/05/age-ubiquitous-screens-some-exhuasted-parents-have-stopped-policing`.

[26] R. H. Thaler and C. R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin, 2009.

[27] B. Ur, J. Jung, and S. Schechter. Intruders versus intrusiveness: Teens' and parents' perspectives on home-entryway surveillance. In *Proc. UbiComp*, 2014.

[28] A. Vasalou, A.-M. Oostveen, and A. N. Joinson. A case study of non-adoption: the values of location tracking in the family. In *Proc. CSCW*, 2012.

[29] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh. A field trial of privacy nudges for facebook. In *Proc. CHI*, 2014.

[30] S. D. Warren and L. Brandeis. The right to privacy. *Harvard Law Review*, IV, December 1890.

[31] P. J. Wisniewski, H. Xu, M. B. Rosson, and J. M. Carroll. Adolescent online safety: The moral of the story. In *Proc. CSCW*, 2014.

[32] S. Yardi and A. Bruckman. Social and technical challenges in parenting teens' social media use. In *Proc. CHI*, 2011.

APPENDIX

A. TEEN INTERVIEW SCRIPT

Good {morning/afternoon}. My name is ___ and my colleague's name is ___. We will be moderating your interview today. Can we get you a glass of water or anything else to drink?

To begin, we would like you and your parent to review this consent form. It contains important information about today's interview. If you and your parent consent to the terms and would like to participate in the study, please sign the form and hand it back to us. [Present consent form]

At this point, we would like to ask your parent to leave. [Addressing parent] Our interview will take approximately one hour. You are welcome to wait outside or return once we are done. [Wait for parent to leave before continuing].

In this research study, we are interviewing a series of teenagers and a separate series of parents of teenagers to understand whether teens have a right to privacy, as well as what that means. We are also trying to understand how parents and teenagers make decisions about using new devices, apps, and websites. As part of this study, we will be asking you questions that relate to your relationship with members of your family. You are free to choose not to answer any questions, or to stop the interview at any point if you feel uncomfortable. We greatly value your honest and candid responses.

We would like to make an audio recording of this session. The members of your family will not listen to this interview recording, and we will not discuss with them what you say during the interview. This recording will only be used for the purposes of this study and will only be accessible to the researchers. Do you consent to having this session audio recorded?

Demographics

1. How old are you?
2. How many people other than you live in your house? What is each person's relationship to you?
3. What grade are you in school?

Online privacy

1. Do you have a computer? Where are the computers located in your house?
   (a) Where are you allowed to use your computer?
   (b) At what age were you first allowed to use a computer?
   (c) Do your parents have the password to your computer?
   (d) Do your parents use the password to check your computer?
   (e) Do your parents monitor your computer use in any other way?
2. Do you have a phone?
   (a) Is it a smartphone?
   (b) Where are you allowed to use your phone?
   (c) At what age were you first allowed to use a phone?
   (d) Do your parents have the password to your phone?
   (e) Do your parents use the password to check your phone?
   (f) Do your parents monitor your phone use in any other way?
3. Do you have a tablet?
   (a) Where are you allowed to use a tablet?
   (b) At what age were you first allowed to use a tablet?
   (c) Do your parents have the password to your tablet?
   (d) Do your parents use the password to check your tablet?
   (e) Do your parents monitor your tablet use in any other way?
4. Do you have a gaming device, like an Xbox or Playstation?
   (a) Where are you allowed to use a game console?
   (b) At what age were you first allowed to use a game console?
   (c) Do your parents monitor your game console use?
5. Do you have an email address? At what age did you sign up for it? Do your parents have the password to this email account? Do your parents monitor your email account in any other way?
6. Do you have a Facebook, Instagram, Twitter, or other social media account? At what age did you sign up for it? Do your parents have the password to this account? Do your parents monitor this account any other way?
7. Do you feel your parents' restrictions are adequate, too much, or too little? Do you feel your parents respect your privacy online? Have you ever tried to get around their restrictions? What changes, if any, would you make to your parents' rules?

Space and New Technologies

1. What do you consider to be **your** space online, where you feel comfortable? What parts, if any, would you be comfortable with your parents seeing? Ideally, what would be your space online?
2. What devices do you own that you would consider your space?
3. What does it mean for something to be your space?
4. Do you look at things online that you wouldn't want your parents to know about? Do you think that your parents might know anyway?
5. How do you hear about new websites and apps?
6. How do you decide whether to join or use these new sites and apps?
7. Do you consider privacy when joining or using them? If so, how do you evaluate the privacy risks?
8. Do you think about your privacy online in the same way as privacy in the physical world, or differently?
9. Do you think your parents understand your privacy needs? Do you think your parents understand what it's like to grow up today, and how it differs from when they grew up?

Privacy at home

1. At home, do you have your own bedroom? If not, with whom do you share your room?
2. Are you allowed to keep the door to your room closed? Why or why not?
3. Do your parents knock before entering your room? Why or why not? At what age did they start knocking?
4. Under what circumstances do you consider it appropriate for your parents to enter your room when you are not there? When is it not appropriate? Are there places within your room that are not appropriate for them to go?
5. Do the bedroom doors in your house have locks? Do you use these locks? If so, when do you use it? Why? Is it appropriate for your parents to unlock your door?
6. Do you feel that your parents give you enough personal time and space at home? As far as you know, do they think they give you enough time and space?

Social privacy

1. What proportion of your friends do your parents know? Do you think you should have to tell your parents about all of your friends?
2. Do your parents impose any restrictions on you going out with friends, such as based on time, people, or location? Do they require you to notify them about where you are going, and with whom? Have you ever broken these restrictions?
3. Are there any rules in your family about dating? In general, are your parents aware of your romantic or sexual experiences?
4. Do your parents give you too little, too much, or just the right amount of space for your social life? How do you feel about these restrictions?

Other

1. Are you aware of any laws relating to children and privacy? What laws do you think there should be?
2. (Optional) Should existing privacy laws be removed or changed?
3. What kinds of information about you, if any, would you not want your parents to share with others? (e.g. family)
4. In general, is it ethical for a parent to look through their teenagers' text messages, Facebook, or email? Are there any circumstances under which your answer would change?
   - Do teenagers have the right not to reveal information to a parent?
   - Do teenagers have the right not to tell their parents about their grades in school?
   - Do teenagers have the right not to tell their parents about health information?
   - In general, do you think that teenagers have a right to privacy from their parents?
   - Do you feel that your parents respect your privacy at home?
   - Are there any other privacy rights which a teenager should or should not have from parents that we have not discussed today?
5. Do you think your siblings' answers to the questions today would have been similar to or different from yours?
6. Do you think your parents would be surprised to hear any of your responses today?
7. Do you have any other comments or questions about any topics we covered today?

Thank you very much for your participation! Your feedback has been valuable to our research.

We will eventually write a research paper about the conversations we have had with you and other research participants. In the paper, we would like to include quotations from some of our participants with attribution in the form of "Participant #." Do you give us permission to use excerpts from this interview in this research paper? Is there anything that we discussed today which you would like us not to quote? Thanks again! [Compensate participant]

B. PARENT INTERVIEW SCRIPT

Good {morning/afternoon}. My name is ___ and my colleague's name is ___. We will be moderating your interview today. Can we get you a glass of water or anything else to drink?

To begin, we would like you to review this consent form. It contains important information about today's interview. If you consent to the terms and would like to participate in the study, please sign the form and hand it back to us. [Present consent form]

In this research study, we are interviewing a series of teenagers and a separate series of parents of teenagers to investigate whether teens have a right to privacy, as well as what that means. We are also trying to understand how parents and teenagers make decisions about using new devices, apps, and websites. As part of this study, we will be asking you questions that relate to your relationship with members of your family. You are free to choose not to answer any questions, or to stop the interview at any point if you feel uncomfortable. We greatly value your honest and candid responses.

We would like to make an audio recording of this session. Please note that the members of your family will not listen to this interview recording, and we will not discuss with them what you say during the interview. This recording will only be used for the purposes of this study and will only be accessible to the researchers and transcribers. Do you consent to having this session audio recorded?

Demographics

1. How many people other than you live in your house? What is each person's relationship to you? Do you have any children who don't live with you?
2. How old are your children, and what grades are they in?

Online privacy

1. Where are the computers located in your house? Does your child have his/her own computer?
   (a) Where is your child allowed to use a computer?
   (b) At what age was your child first allowed to use a computer?
   (c) Do you have the password to your child's computer?
   (d) Do you use the password to check your child's computer?
   (e) Do you monitor your child's computer use in any other way?
2. Does your child have a phone?
   (a) Is it a smartphone?
   (b) Where is your child allowed to use his/her phone?
   (c) At what age was your child first allowed to have his/her own phone?
   (d) Do you have the password to your child's phone?
   (e) Do you use the password to check your child's phone?
   (f) Do you monitor your child's phone use in any other way?
3. Does your child have a tablet, like an iPad?
   (a) Where is your child allowed to use a tablet?
   (b) At what age was your child first allowed to use a tablet?
   (c) Do you have the password to your child's tablet?
   (d) Do you use the password to check your child's tablet?
   (e) Do you monitor your child's tablet use in any other way?
4. Does your child have a gaming device, like an Xbox or Playstation?
   (a) Where is your child allowed to use a gaming device?
   (b) At what age was your child first allowed to use a gaming device?
   (c) Do you monitor your child's gaming use? How?
5. Does your child have an email address? At what age did they sign up for it? Do you have the password to this email account? Do you monitor this email account any other way?
6. Does your child have a Facebook or other social media account? At what age did they sign up for it? Do you have the password to the account? Do you monitor this account any other way? (Are you friends with them?)
7. Do you feel your restrictions are adequate, too much, or too little? Do you feel your child has the right amount of personal space online? Do you suspect your child has ever tried to hide their online activity from you? What changes, if any, would you consider making to your rules?

New technologies

1. How do you hear about new devices, websites, and apps that teenagers are using these days? What about technologies your children use themselves?
2. How do you decide what rules, policies, and strategies to adopt regarding your teen's use of these devices, websites, and apps?
3. Do you have any concerns about your teen's privacy with new devices, websites, and apps?
4. How do you evaluate the privacy risks of new devices, websites, and apps?
5. Do you think about your teen's privacy online in the same way as privacy in the physical world, or differently?

Privacy at home

1. At home, does your child have their own bedroom? If not, with whom do they share the room?
2. Is your child allowed to keep the door to their room closed? Why or why not?
3. Do you knock before entering your child's room? Why or why not? At what age did you start knocking?
4. Under what circumstances do you consider it appropriate to enter your child's bedroom when they are not there? When is it not appropriate? Are there places within their room that are not appropriate for you to go?
5. Do the bedroom doors in your house have locks? Does your child use the locks? When is it appropriate for them to do so? When is it not appropriate?
6. Do you feel that you give your child enough personal time and space at home? As far as you know, does your child think you give them enough time and space?

Social Privacy

1. What proportion of your child's friends do you feel you know? Would you be surprised if your child has friends you are not aware of?
2. Do you impose any restrictions on your child going out with friends, such as based on time, people, or location? Do you require your child to notify you about where he/she is going, and with whom? Do you suspect your child has ever broken these restrictions?
3. Are there any rules in your family about dating? Do you feel you are aware of your child's romantic or sexual experiences?
4. In general, how well does your child keep you informed about his/her life? Are there things you wish he/she would tell you more about?
5. Do you feel your restrictions are adequate? Do you feel you give your child enough, too much, or just the right amount of space for their own social lives? How do you think your child feels about these restrictions?

Other

1. Is there anything your child wouldn't want you to share with others? What kinds of information about your child, if any, would you not share with immediate family members? Extended family members? Friends?
2. In general, is it ethical for a parent to look through their teenagers' text messages? What about their Facebook? Email? Are there any circumstances under which your answer would change?
   - Do teenagers have the right not to reveal information to a parent?
   - Do teenagers have the right not to tell their parents about their grades in school?
   - Do teenagers have the right not to tell their parents about health information?
   - In general, do you think that teenagers have a right to privacy from their parents?
   - Do you feel that you respect your child's privacy at home?
   - Are there any other privacy rights which a teenager should or should not have from parents that we have not discussed today?
   - When you were your son's/daughter's age, did you feel that your parents respected your privacy? Why or why not?
   - Do you feel your view of teen privacy is different from how your parents viewed it when you were a teenager?
3. Do you have any other comments or questions about any topics we covered today?

Thank you very much for your participation! Your feedback has been valuable to our research. [Compensate participant]

Privacy Attitudes of Mechanical Turk Workers and the U.S. Public

Ruogu Kang[1], Stephanie Brown[1,2], Laura Dabbish[1], Sara Kiesler[1]

HCI Institute[1]
Carnegie Mellon University
Pittsburgh, PA
{ruoguk, dabbish, kiesler}@cs.cmu.edu

School of Communication[2]
American University
Washington, DC
sb9279a@student.american.edu

ABSTRACT

Amazon Mechanical Turk (MTurk) is a crowdsourcing platform widely used to conduct behavioral research, including studies of online privacy and security. We studied how well the privacy attitudes of MTurk workers mirror the privacy attitudes of the larger user population. We report results from an MTurk survey of attitudes about managing one's personal information online and policy preferences about anonymity. We compare these attitudes with those of a representative U.S. adult sample drawn from a separate survey a few months earlier. MTurk respondents were younger and better educated, and more likely to use social media than the representative US adult sample. Although they reported a similar amount of personal information online, U.S. MTurk workers put a higher value on anonymity and hiding information, were more likely to do so, had more privacy concerns than the larger U.S. public. Indian MTurk workers were much less concerned than American workers about their privacy and more tolerant of government monitoring. Our analyses show that these findings hold even when controlling for age, education, gender, and social media use. Our findings suggest that privacy studies using MTurk need to account for differences between MTurk samples and the general population.

1. INTRODUCTION

Amazon Mechanical Turk (www.mturk.com) is an increasingly popular platform for conducting behavioral research. It is now widely adopted by researchers in many domains, including psychology [9], economics [18,39], and political science [35]. It is broadly recognized as a fast and inexpensive way to collect data requiring human participation, and provides a level of cultural diversity hard to obtain with other recruitment methods [10,13,35]. MTurk has also become a valuable resource for privacy and security research and is widely used to survey people's opinions on privacy-related issues [5,23,28,29,46]. Researchers have conducted experiments on MTurk to study the effects of framing on information disclosure [6] and the factors influencing people's attitudes toward online behavioral advertising [28]. Others have implemented surveys on MTurk to study users' privacy preferences for mobile apps [29], their privacy concerns on social networking sites [46], and their attitudes about national security [33]. However, none of the previous work has compared the privacy experiences and opinions of MTurk workers with those of the general public. We do not yet know whether privacy research conducted on MTurk is generalizable to other populations.

We address in this paper the comparability of MTurk worker privacy attitudes and behavior with those of the general population. MTurk workers, as with any self-selected subset of the population, may differ from the general population and these differences can constrain the generalizability of study results. One reason to expect differences in their responses is that the privacy practices, social norms, and default settings of different websites may attract different types of people. MTurk's policy is that "collecting personal identifiable information" is prohibited when requesters recruit workers from the market [2]. Thus it may attract people who particularly value privacy. By contrast, the social networking site Facebook encourages real-name accounts, perhaps attracting people who desire, or at least do not oppose, being known. In addition, most workers on MTurk come from two culturally different countries: the U.S. and India. The two countries' different government policies and cultural backgrounds may strongly affect people's experiences and attitudes, which bring additional challenges to privacy research conducted on MTurk.

Our goal is to contribute to the research in online privacy and security by comparing the privacy attitudes of MTurk workers, assessed online, with those of a representative sample of the U.S. public. Our purpose is to understand how the former group's attitudes reflect or diverge from those of the general public. Our results can help researchers calibrate their findings from MTurk samples, and understand their generalizability to broader samples of the public.

We report here two comparisons--a comparison of a U.S. based MTurk worker sample with a representative telephone sample of the U.S. public that uses the Internet, and a comparison of the same U.S. MTurk sample with a sample of Indian MTurk workers. We studied their comparability with respect to two topics: (1) how they manage their personal information online, and (2) their attitudes and preferences regarding privacy and anonymity online. We tackled these topics because the Internet increasingly exposes personal information about people, not just to their intended audiences, but also to third parties who may be completely hidden to them [38]. Recent news events imply that it is difficult if not impossible to put walls around one's content by communicating anonymously or securing access from unauthorized others [1,3]. Even MTurk, an anonymous platform, may fail to protect personally identifiable information about some

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9-11, 2014, Menlo Park, CA.

of its workers [27].

2. RELATED WORK

The rise of greater Internet transparency and threats to personal information has prompted considerable research on what Internet users know about their personal information online and how they try, or fail to try, to protect it [e.g., 21]. National polls suggest that many Internet users do not know how much of their personal information is online and open to surveillance from people they have not authorized to see or use it [29]. Many Internet users have only a vague notion of how the Internet works [36] and the potential threats to their privacy [20]; most do not know who has access to information about them or how people get this information [24]. And the public has become more worried. Kang et al.'s [22] interview study revealed a variety of real-life circumstances and goals that led interviewees to seek anonymity or to hide identified content from particular individuals or organizations.

We began this work with the hypothesis that MTurk workers may have more concerns about privacy than the average member of the Internet-using public. First, these workers are a self-selected group that has chosen an anonymous worksite. Second, recent studies comparing MTurk with other samples suggest that MTurk workers are better educated, more liberal, and younger than the general population [35]. In domains other than privacy, researchers have compared MTurk samples with other groups such as representative samples of the U.S. public, examining differences in their demographic characteristics [10,19,33,35,42], personality traits [7,15], and political attitudes and responses to experimental treatments [8]. Berensky, Huber, and Lenz [8] found that MTurk workers are more representative than convenience samples (e.g., locally-recruited students) but less representative than Internet panels or national representative samples. Goodman, et al. [14] argued MTurk sample was not different from a community sample in their demographics except that MTurk has more international participants. Ross, et al. [42] found an increasing proportion of young people, males, and people with lower income in active Turkers. They also found Indian workers on MTurk to be younger than U.S. workers, and have lower incomes and higher education levels. These differences might predispose MTurk workers to be more knowledgeable about threats to online privacy. Martin et al. [32] studied a crowd workers' forum and found that MTurk workers do desire anonymity and tend to avoid surveillance. Lease's survey [26] of 1,000 MTurk workers suggests that they place a high value on the anonymity of their names and home addresses. This prior work raises the possibility that MTurk workers may have a higher level of concern than the general Internet-using public about the privacy and protection of their personal information.

The majority of MTurk workers are from the U.S. and India [10,13,35]. The proportion of U.S. workers and Indian workers on MTurk in a recent study [42] was 57% and 32% among the 573 workers openly recruited. Collecting data on MTurk can help researchers investigate cultural differences, but at the same time might introduce extra noise into their data because of geographic and cultural differences. People's privacy perceptions and preferences are often shaped by the societies in which they live and by their cultural backgrounds [45]. From prior work, we expected differences in privacy attitudes of U.S. MTurk workers and Indian MTurk workers. Americans disclose more in traditional communication settings than people from Eastern cultures (Chinese in [11]). Recent work suggests, however, that Americans disclose less online than face-to-face because they have more concerns about online communication being exposed to others [47]. Indians seem to have a lower degree of privacy concern than Americans [25,45,46], and place more trust in government organizations that collect personal information [25]. Mason and Dupuis [33] compared Indian and U.S. MTurk workers' perceptions about security and their opinions about Edward Snowden's revelations of NSA surveillance. Compared with U.S. MTurk workers, Indians reported greater agreement with statements such as "Snowden's actions have damaged U.S. national security," suggesting that the American workers were more inclined to value Snowden's revelations. Therefore, another purpose of this work was to compare the privacy attitudes and behavior of a U.S. MTurk worker with an Indian MTurk worker sample.

In sum, our research aims to answer two questions: (1) Are U.S. MTurk workers different from the U.S. Internet users in managing their personal information and in their opinions about online privacy? (2) Are U.S. MTurk workers different from Indian MTurk workers in managing their personal information and in their opinions about online privacy?

3. METHOD

3.1 Participants and procedure

We compared responses in two survey studies of privacy and anonymity, one a representative telephone sample of U.S. Internet users and the other (a few months later) an online survey of MTurk workers. Most items for both surveys were the same. We report here only the responses to questions that were identically- or similarly-phrased on both surveys. Because the surveys given to the representative U.S. sample were conducted by phone with voice responses, and the MTurk surveys were conducted online, with typed responses, the response options were never identical. However, as much as feasible, the questions themselves were identical. The survey questions we analyzed in this paper are attached in an appendix.

The first survey was administered by the Pew Research Center's Internet Project (referred to here as "Pew") in July 11-14, 2013. The current authors collaborated with Pew researchers on constructing questions for this survey. The survey items were developed based on the interview questions about anonymity in Kang et al. [22] and questions on privacy that the Pew Research Center fielded in its previous surveys [30,31,37]. Pew surveyed a representative sample of U.S. adults consisting of 1,002 U.S. adults ages 18 and over, with 500 surveys using landline telephones and 500 surveys using cell phones. Respondents were not paid, except any cell phone charges were reimbursed. When conducting the survey, interviewers asked respondents if they would be willing to participate in a confidential and anonymous survey. Participants were then asked a series of questions, first to determine if they were Internet users, and then about their activities online. Of the total participants, 775 said they used the Internet and our analysis is based on responses from these Internet users.

The authors conducted the second survey on Amazon Mechanical Turk. We recruited 418 people from MTurk from February 16-20, 2014. We used the same sampling criteria as in previous studies to increase quality [23, 35], by restricting the participants to those with an approval rate of at least 95% and at least 100 approved HITs. Each participant was paid $1 for completing the survey.

They were told that the survey was about how people use the Internet. Separate HITs were released to recruit participants from the U.S., India and other countries. After accepting the HIT, MTurk workers were directed to a SurveyMonkey survey. The survey was completely voluntary and confidential. Participants could opt out of the survey at any time. Twenty-two responses (5%) were excluded because they failed the attention check questions or entered invalid responses. The dataset we analyze here includes 310 valid responses: 182 from the U.S. and 128 from India.

3.2 Survey items

The Pew survey and MTurk survey posed questions related to privacy and anonymity, and demographics.

3.2.1 Managing their personal information

Both surveys asked respondents to estimate what personal information is online for others to see: "Is any of the following information about you available on the Internet for others to see? It doesn't matter if you put it there yourself or someone else did so." They were asked about their email address, home address, home phone number, cell phone number, employer or company, political party or political affiliation, things they've written with their name on it, photo of themselves, video of themselves, which groups/organizations they belong to, and birth date.

Both surveys also asked respondents whether they had tried to hide their identity online: "Have you ever tried to use the Internet in a way that hides or masks your identity from certain people or organizations?" Those who answered "yes" to this question were coded as having tried to hide their identity.

Internet users may be differently concerned about protecting their personal information when they communicate with different groups. To study whether respondents were selective in hiding content (such as posts) that they had communicated online, the national sample Pew survey asked participants "Have you ever tried to use the Internet in ways that keep ___ from being able to see what you have read, watched or posted online?" They were asked if they had done this to "family members or a romantic partner;" "certain friends;" "people from your past;" "an employer, supervisor, or coworkers;" "the companies or people who run the website you visited;" "hackers or criminals;" "law enforcement;" "people who might criticize, harass, or target you;" "companies or people that might want payment for the files you download such as songs, movies, or games;" "advertisers;" "the government?" In the MTurk survey, we slightly modified the format and combined the responses in our analysis in order to compare them with the national sample. (The different ways in which the questions were asked may matter, so this comparison should be considered only in context of the whole.)

3.2.2 Privacy attitudes and preferences

Both surveys asked respondents, "Do you ever worry about how much information is available about you on the Internet?" The respondents also were asked about their opinions about government policies: "Do you think the laws provide reasonable protections of people's privacy about their online activities?" and their opinions about anonymity: "Considering everything you know and have heard about the Internet, do you think it is possible for someone to use the Internet completely anonymously – so that none of their online activities can be easily traced back to them?" and "Do you think that people should have the ability to use the Internet completely anonymously for certain kinds of online activities?"

We do not report on additional questions about privacy that were not similar in the two surveys.

4. RESULTS

We used JMP statistical software to conduct multivariate and regression analyses of the survey data. Only a few respondents opted not to answer particular questions so we did not need to adjust for missing data.

4.1 Demographic differences

We first compare the demographic characteristics of the U.S. public sample, the U.S. MTurk sample, and the Indian MTurk sample (Table 1). Our MTurk samples seem similar to MTurk samples in other studies, for instance the 2,912 participants in [28]. Consistent with previous studies [8,23], our MTurk samples are younger and the Indian sample is better educated than the U.S. public sample (81% have a college education or higher, compared with the U.S. MTurk sample, t [308] = 6.29, $p < .01$). Both MTurk samples had more male than female respondents, whereas the U.S. public representative sample had equal male and female respondents. The MTurk samples are also much more likely to use social media. Because social media tends to elicit personal information from people, using social media should predict more

Demographic Characteristics	**U.S. Public**	**U.S. Turk**	**Indian Turk**
N	**775**	**182**	**128**
Age			
18-24	12%	24%	23%
25-34	14%	41%	56%
35-44	13%	23%	12%
45-54	17%	9%	5%
55-64	24%	3%	2%
65+	19%	1%	2%
Mean age	49.8	32.7	30.5
	F [2,1080] = 122.72, $p < .001$		
Gender			
Female	50%	42%	35%
Male	50%	57%	65%
	X^2 [2, 1084] = 11.76, $p < .01$		
Education			
High school or less	26%	12%	5%
Some college	31%	45%	14%
College and more	42%	43%	81%
	F [2,1080] = 24.62, $p < .001$		
Percent who use social media	68%	90%	98%
	X^2 [2,1085] = 97.04, $p < .001$		

Table 1. Demographic characteristics of three datasets: U.S. telephone representative sample (referred to as U.S. public in paper), U.S. Turk sample and Indian Turk sample. Total N = 1085.

concerns about privacy as well [40].

The demographic characteristics of a group of people may be highly predictive of their attitudes. For instance, younger people may be more politically liberal (among many other differences) than older people [8]. This expectation leads us to ask whether any difference in privacy attitudes between MTurk workers and the U.S. general public might be due merely to their being younger, for example, rather than to their being MTurk workers. That is, would any younger group respond the same way as the MTurk sample? To tackle this question, we conducted hierarchical regression analyses. For each of the dependent variables, we first conducted a regression analysis using a dummy variable of the two samples as a predictor variable (Model 1), then added age to the model (Model 2), and finally, we added gender, education, and social media use to the regression model (Model 3). If demographic variables explain differences in privacy attitudes, then these factors should contribute a statistically significant effect, and the effect of the U.S. MTurk vs. U.S. public samples should become less significant or insignificant.

4.2 Comparing the U.S. Internet-using public with U.S. MTurk workers

In Table 2, we show the results of our comparisons between the U.S. public and U.S. MTurk workers and the hierarchical regressions.

4.2.1 Managing their personal information

The first row compares how much personal information the U.S.

		Independent variables					
Dependent variables	**Model**	**Sample (U.S. Turk=1)**	**Age**	**Gender (Male=1)**	**Education**	**Use social media**	**R^2**
Managing their personal information							
Amount of available information online (Above median number of items of information = 1)	1	1.25 (.90, 1.74)					.002
	2	.89 (.62, 1.26)	.98*** (.97, .99)				.032
	3	.71 (.49, 1.03)	.98*** (.98, .99)	1.06 (.80,1.39)	1.21***(1.11, 1.31)	3.66*** (2.64, 5.12)	.113
Hide identity (Yes = 1)	1	2.23***(1.52,3.23)					.017
	2	1.79**(1.19, 2.67)	.99**(.98, .99)				.025
	3	1.58*(1.05, 2.37)	.99*(.98, .99)	1.30 (.93,1.83)	1.18***(1.07, 1.31)	2.58***(1.60, 4.33)	.059
Hide online content from people or organizations (Hide content from at least one group = 1)	1	2.37***(1.67,3.42)					.025
	2	1.69**(1.16, 2.48)	.98***(.97, .99)				.054
	3	1.47*(1.00,2.17)	.98***(.98, .99)	1.20 (.91,1.58)	1.11*(1.02, 1.20)	2.79***(2.03, 3.85)	.101
Privacy attitudes and preferences							
Worry about information available on the Internet (Yes=1)	1	1.67**(1.20, 2.35)					.009
	2	1.66**(1.17, 2.38)	1.00(.99, 1.01)				.009
	3	1.55*(1.09, 2.24)	1.00(.99, 1.01)	.80 (.61, 1.04)	1.08* (1.01,1.17)	1.38* (1.01, 1.88)	.022
Think that the laws provide reasonable protections of people's privacy (Yes = 1)	1	.73 (.47, 1.09)					.003
	2	.66 (.42, 1.01)	.99 (.99, 1.00)				.005
	3	.67 (.43, 1.03)	1.00 (.99, 1.01)	1.03 (.75, 1.41)	.94 (.86, 1.03)	1.17 (.80, 1.72)	.007
Think that it is possible to be completely anonymous (Yes=1)	1	.76 (.52, 1.09)					.002
	2	.69 (.46, 1.01)	.99 (.99, 1.00)				.005
	3	.71 (.48, 1.06)	1.00 (.99, 1.01)	1.58**(1.19,2.10)	.87**(.80, .95)	1.14 (.82, 1.60)	.028
Think that people should have the ability to be anonymous online (Yes = 1)	1	3.63***(2.31,5.98)					.039
	2	2.67***(1.66,4.48)	.98***(.97, .99)				.060
	3	2.55***(1.58,4.30)	.98***(.97, .99)	1.73***(1.29,2.33)	1.02 (.93, 1.11)	1.18 (.84, 1.65)	.075

*** p < .001, ** p < .01, * p < .05. Values in the table are odds ratios (95% CI). An odds ratio that is larger than 1.0 indicates positive prediction, and an odds ratio that is smaller than 1.0 indicates negative prediction. If the 95% confidence interval for an odds ratio does not contain 1.0, the association is statistically significant at .05 level. N = 957.

Table 2. Hierarchical logistic regression showing the effects of sample differences (U.S. Turk vs. U.S. public), demographic variables, and social media use on privacy behavior and attitudes.

MTurk sample reported having online as compared with the same report of the U.S. public sample. For simplicity of presentation in Table 2, we show how many items of information (e.g., phone numbers, address, photos of self) are above vs. below the overall median number of items reported in both samples (median number = 4). We did not find statistically significant differences (Mean of U.S. MTurk = 4.2, Mean of U.S. public = 3.9, t [951] = 1.35, p = .18), meaning that the two samples did not differ in how many items of information they had online (see Figure 1a, red and green lines). Model 2 adds an estimate of the effect of age and Model 3 adds the additional effects of gender, education, and social media. These results show that younger respondents, those with more education, and those who use social media reported having more personal information online than their counterparts. These findings confirm the important predictive value of demographic factors.

The next row of findings in Table 2 looks at the question "Have you ever tried to use the Internet in a way that hides or masks your identity from certain people or organizations?" We found that U.S. MTurk workers were significantly more likely to seek anonymity than the U.S. public generally (31% vs. 17%, t [939] = 4.30, p < .001). This difference remained significant when we added age (Model 2) and (education, gender, and social media use) into the prediction (Model 3). Thus, we found that younger people, people with higher education levels, and people who use social media were more likely to have ever sought anonymity or hid their identity but even controlling for these factors, MTurk workers were also more likely to have done so (see Figure 1b red and green lines).

Pseudonyms are considered an important method of protecting one's privacy [45]. The U.S. public survey asked respondents if they had posted online using their real names, usernames associated with their true identities, or without revealing who they are. Thirty-three percent of the U.S. public sample said they had posted without revealing who they are. In the MTurk survey, the question was different (therefore not shown in Table 2). We asked respondents if they ever posted using a username that people did not associate with them, and if they posted using no name at all. Eighty-one percent of the U.S. MTurk respondents said "yes" to at least one of these last two choices. Although these questions are not the same across the two samples, the results combined with those in Table 2 suggest that U.S. MTurk workers may attempt to use unidentifiable communications or hide their identity more than the U.S. public.

The third row in Table 2 shows whether respondents try to hide their online contributions or content selectively, from different groups such as friends or employers. Significantly more participants in the U.S. MTurk sample reported having tried to hide content from at least one group than in the U.S. public sample (73% vs. 53%, t [955] = 4.94, p < .001). This difference remained even when adding demographic variables into the regressions. The percent of people who had hidden content from at least one group in the two samples is shown in Figure 1c red and green lines. In delving into this question more specifically, we found that U.S. MTurk workers had tried to hide content from their family members, a romantic partner, certain friends, or coworkers than U.S. public had (54.4% vs. 19.3%, t [954] = 10.24, p < .001); the same is true for their employers, supervisors or companies they work for (26.9% vs. 9.8%, t [926] = 6.26, p < .001); and for law enforcement, government, or companies or people that may want payment for the files that they downloaded (18.1% vs. 10.5%, t [952] = 2.87, p < .01). However, respondents in the U.S. public sample were significantly more likely to report hiding from hackers, criminals, or advertisers than the U.S. MTurk workers (43.6% vs. 28%, t [948] = 3.88, p < .001). The two samples did not show any significant difference in hiding content from people from the past and people who might criticize, harass or target them.

The analyses of the effects of demographic variables showed a similar pattern as the prior question about hiding one's identity:

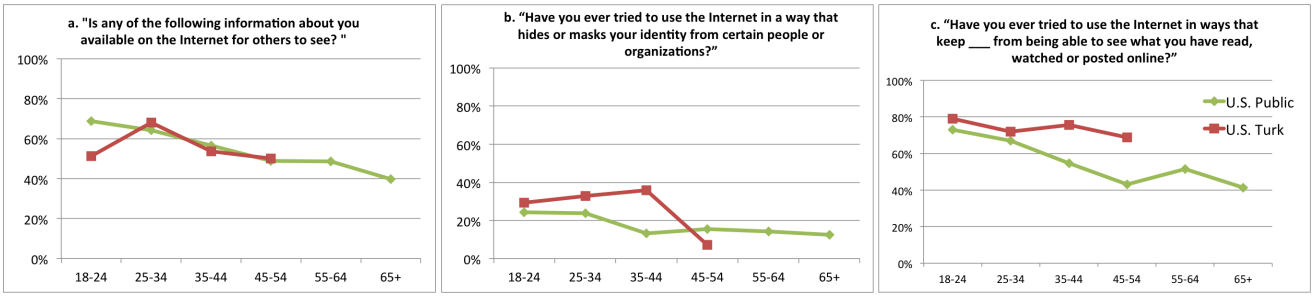


Figure 1. Percent of respondents who said yes to three questions about how they manage their personal information. (MTurk data for those over age 55 excluded due to the few number of respondents in these samples.) Note. The data shown in figure 1a is the percent of people who reported more than the median number of items online.

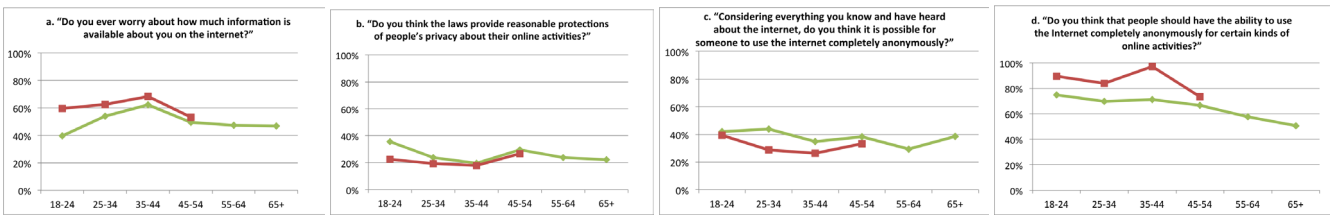


Figure 2. Percent of people who answered yes to each of four privacy preferences questions. (MTurk data for those over age 55 excluded due to the few number of respondents in these samples.)

younger, more educated respondents, and social media users (but not men or women) were more likely to protect their personal information from other people or groups.

4.2.2 Privacy attitudes and preferences

Are U.S. MTurk workers more concerned about privacy than the general U.S. public? Our results suggest the answer is yes. Table 2 shows how the two samples differ in their privacy preferences and concerns. U.S. MTurk workers in our study expressed more concern about their information than the U.S. public. Sixty-three percent of the U.S. MTurk workers said they worried about how much information is available about them on the Internet, while only 50% of the U.S. public sample said this (t [948] = 3.04, $p <$ .01). Adding demographic variables and social media use in the models, the effect of the sample difference dropped only slightly and remained significant. This finding suggests that U.S. MTurk workers are more worried about their online information than the U.S. public, regardless of their age, gender, education, and social media use. Additionally, there is a separate effect of education level and social media predicting these concerns. Those with higher education and those who use social media are more likely to worry about their personal information online. Figure 2a shows the percent of people who worry about their information in different age groups.

We were also interested in people's policy preferences. Our analysis showed that U.S. MTurk workers did not differ significantly from the U.S. public in their opinions about whether current privacy laws provide enough protection of their privacy (Figure 2b). Only eighteen percent of the U.S. MTurk workers thought current laws provide reasonable protection of people's privacy, and 23% of the Pew sample said so. None of the demographic variables and the social media use made a difference either.

Prior work suggests most people, regardless of nationality or experience, understand that anonymity has tradeoffs. They believe that anonymity can be misused and can encourage irresponsible behavior without consequences for the perpetrators [22]. And there is evidence that anonymity can encourage negative social behavior [12,44]. On the other hand, anonymity might help people avoid negative online experiences and persons or groups from whom they wish to hide [22]. We wanted to know whether respondents thought anonymity is possible on today's Internet and whether they should have the ability to be anonymous online. We asked: "Considering everything you know and have heard about the Internet, do you think it is possible for someone to use the Internet completely anonymously – so that none of their online activities can be easily traced back to them?" We found that 37% of the U.S. public respondents and 31% of the U.S. Turk sample thought that it was possible to be completely anonymous online and there was no significant difference between the two samples. Male and lower education respondents agreed more strongly anonymity is possible. We also asked, "Do you think that people should have the ability to use the Internet completely anonymously for certain kinds of online activities?" Our results showed that anonymity is embraced among more U.S. MTurk workers (Figure 2c). The percentage of the U.S. MTurk sample who said people should have the ability to be anonymous online was significantly higher than in the U.S. public sample (86% vs. 63%, t [883] = 5.74, $p <$ .01). The difference between the two samples remains significant when we add more demographic information into the model (Models 2 and 3). Separately, demographic factors predicted people's anonymity preferences: younger people and men preferred more anonymity than their counterparts.

4.2.3 Summary of findings

The results comparing the U.S. MTurk worker and U.S. representative public samples show that on four of seven items, U.S. MTurk workers differed from the U.S. sample, even when demographic variables and social media use were taken into account (Table 2). Although they have the same amount of personal information online, more MTurk workers have tried to be anonymous, they have tried to hide their contributions from more different audiences, are more worried about their online information, and believe they should be able to communicate anonymously online. Their opinion about whether or not it is possible to be completely anonymous online, however, is not significantly different. Another important point is, as shown in Figure 1 and 2, the two samples show similar trends in how their behaviors and attitudes change based on age. Younger people seem to have more personal information online, but also have stronger tendency towards hiding their online identity and content.

4.3 Comparing U.S. MTurk workers with Indian MTurk workers

We analyzed the same set of questions in our survey answered by U.S. MTurk workers and Indian MTurk workers. We conducted analyses shown in Table 3 to compare their responses.

4.3.1 Managing their personal information

On average, Indian MTurk workers reported that more of their personal information was online than U.S. MTurk workers did (M [Indian MTurk workers] = 5.7, M [U.S. MTurk workers] = 4.2; t [308] = 5.35, $p < .001$). The difference between two samples remains significant when we add demographic variables into the model and whether they use social media or not in the model (Models 2 and 3, the first row in Table 3). None of the demographic variables had an effect on their perception of online information, but using social media predicted more personal information online. Figure 3a, blue vs. red lines, shows the comparison between U.S. Turkers and Indian Turkers.

We also found U.S. MTurk workers were more likely to seek to hide their identity than Indian MTurk workers (31% vs. 16%; t [285] = 2.88, $p <$ .01, Figure 3b). As shown in the second row in Table 3, we did not find any significant demographic variables explaining the difference, so we can conclude that, for the variables we have studied, the two groups differ in their anonymity-seeking behavior.

Although more U.S. MTurk workers reported seeking anonymity, they did not name more people or groups they were hiding from than Indians MTurk workers did (73% vs. 76% in each sample named at least one individual or group that they have hidden content from). As shown in Models 2 and 3 in the third row of Table 3, the two samples did not show any difference but younger respondents hid from more groups across both samples (Figure 3c).

When we looked at each sample specifically (Figure 5), we saw that significantly more Indian MTurk workers reported hiding from employers or supervisors than U.S. MTurk workers (42% vs. 27%, t [307] = 2.75, $p <$ .01), and slightly (but not significantly) more Indian MTurk workers hid from people from the past, those who might criticize them, and hackers, criminals, or advertisers

Dependent variables	Model	Independent variables: Sample (U.S. Turk=1)	Age	Gender (Male=1)	Education	Use social media	R^2
Managing their personal information							
Amount of available information online (Above median number of items of information = 1)	1	.37***(.22, .61)					.047
	2	.35***(.20, .59)	.99 (.96, 1.01)				.056
	3	.43**(.24, .75)	.99 (.97, 1.02)	1.58 (.94, 2.67)	1.07 (.90, 1.27)	4.79**(1.81,14.25)	.093
Hide identity (Yes = 1)	1	2.36**(1.32,4.38)					.029
	2	2.41**(1.34,4.51)	.98 (.95, 1.01)				.034
	3	3.26***(1.70,6.53)	.98 (.94, 1.01)	1.51(.85,2.76)	1.21 (.99,1.49)	1.48(.49, 5.50)	.052
Hide online content from people or organizations (Hide content from at least one group = 1)	1	.86 (.51, 1.44)					.001
	2	.97 (.57, 1.65)	.96**(.94, .99)				.031
	3	1.17 (.65, 2.09)	.96**(.94, .99)	1.07 (.62, 1.85)	1.15 (.96, 1.38)	1.60 (.58, 4.16)	.041
Privacy attitudes and preferences							
Worry about information available on the Internet (Yes=1)	1	3.01***(1.85,4.95)					.065
	2	2.90***(1.78,4.78)	1.00 (.98, 1.03)				.062
	3	3.17***(1.87,5.50)	1.00 (.98, 1.03)	.70 (.42, 1.15)	1.08(.92,1.29)	1.37(.53,3.58)	.074
Think that the laws provide reasonable protections of people's privacy (Yes = 1)	1	.19***(.11, .32)					.128
	2	.19***(.11, .32)	1.01 (.98, 1.04)				.125
	3	.17***(.09, .30)	1.02 (.99, 1.05)	1.50(.85, 2.69)	.91 (.74, 1.09)	.84 (.28, 2.82)	.135
Think that it is possible to be completely anonymous (Yes=1)	1	.26***(.15, .43)					.095
	2	.27***(.16, .45)	.99 (.96, 1.02)				.093
	3	.29***(.16, .50)	.99 (.96, 1.02)	1.31 (.75, 2.30)	1.14(.95,1.37)	.43 (.16, 1.16)	.108
Think that people should have the ability to be anonymous online (Yes = 1)	1	1.92*(1.03, 3.6)					.015
	2	1.87*(1.00, 3.54)	1.00 (.97, 1.04)				.014
	3	1.97*(1.00, 3.92)	1.00 (.97, 1.03)	.97 (.49, 1.87)	1.08(.87,1.35)	.60 (.09, 2.30)	.017

*** p < .001, ** p < .01, * p < .05. Values in the table are Odds Ratio (95% CI). Odds ratio that is larger than 1.0 indicates positive prediction, and odds ratio that is smaller than 1.0 indicates negative prediction. If the 95% confidence interval for OR does not contain 1.0, the association is statistically significant at .05 level. N = 310.

Table 3. Hierarchical logistic regression showing the effects of sample differences (U.S. Turk vs. Indian Turk), demographic variables, and social media use on privacy behavior and attitudes.

(35% vs. 27%, t [306] = 1.53, p = .13). Their experiences with the other three groups did not show significant difference.

4.3.2 Privacy attitudes and preferences

Although more of their information was online and more of them used social media, Indian MTurk workers were significantly less worried than U.S. MTurk workers about their personal information on the Internet (the fourth row in Table 3; Figure 4a). Sixty-two percent of the U.S. MTurk workers said they worried about how much information was available about them on the Internet, but only 35% of the Indian participants said this (t [289] = 4.66, $p < .001$). Adding demographic variables and social media use in the model did not reduce the significant effect of the sample difference (Models 2 and 3 vs. 1). The finding suggests that U.S. MTurk workers have more concerns about their personal information online than Indian MTurk workers, regardless of their age, gender, education and whether they use social media or not.

We also found consistent significant differences between Indian and U.S. MTurk workers' policy preferences and their opinions about anonymity. U.S. MTurk workers showed more dissatisfaction about how the government protects their privacy than Indian MTurk workers (the fifth row in Table 3): only 18% of the U.S. MTurk workers said current laws provide reasonable protection of people's privacy, whereas 52% of the Indian participants thought their laws provide enough protection of their privacy (t [281] = 6.95, $p < .001$, Figure 4b). Less U.S. than Indian MTurk workers believed that people could achieve complete anonymity on today's Internet (31% vs. 64%, t [259] = 5.51, $p < .001$, the sixth row in Table 3, Figure 4c). More U.S. than Indian MTurk workers said people should have the ability to use the Internet completely anonymously (86% vs. 77%, t [276] = 2.10, p = .04, the seventh row in Table 3, Figure 4d).

Consistent with this finding, a question added to the MTurk survey (that was not posed in the U.S. public survey) asked respondents whether the government should be able to monitor everyone's email and other online activities "if officials say this might prevent future terrorist attacks." Fifty-seven percent of the Indian MTurk workers agreed with this statement but only 9% of

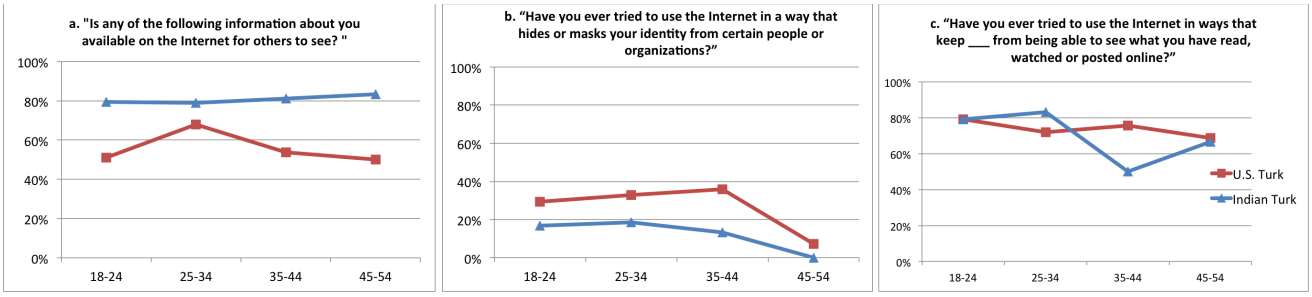


Figure 3. Percent of respondents who said yes to three questions about how they manage their personal information. (MTurk data for those over age 55 excluded due to the few number of respondents in these samples.) Note. The data shown in figure 1a is the percent of people who reported more than the median number of items online.

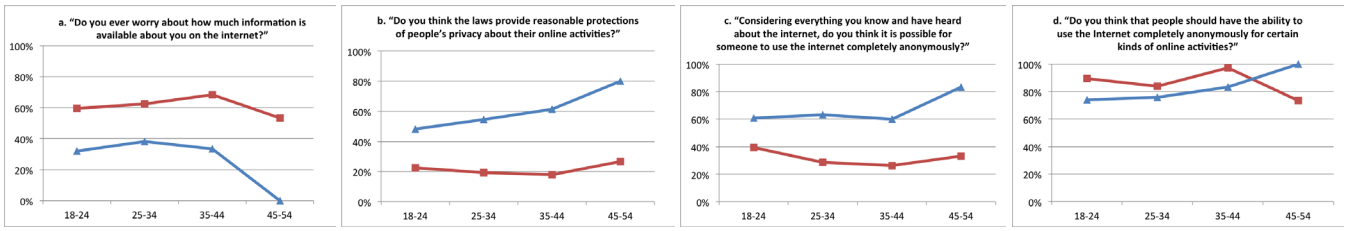


Figure 4. Percent of people who answered yes to each of four privacy preferences questions. (MTurk data for those over age 55 excluded due to the few number of respondents in these samples.)

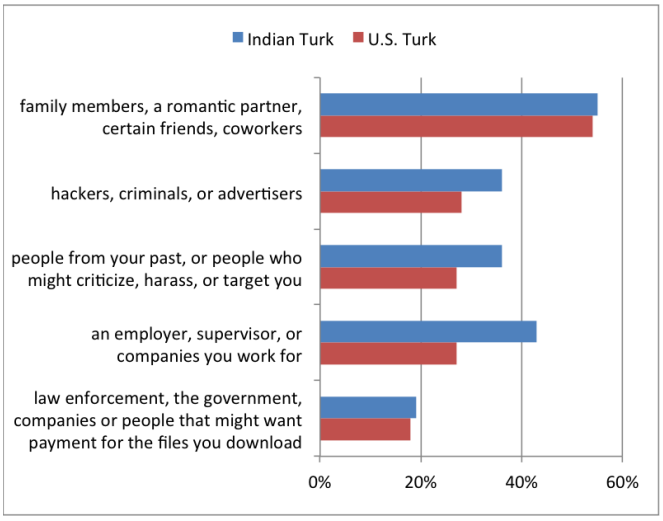


Figure 5. Percent who reported hiding content from certain people or groups.

the U.S. MTurk workers agreed (t [268] = 9.88, p < .001). A different national U.S. survey [38] asking the identical question showed somewhat higher agreement among the U.S. public (45%) as compared to the U.S. MTurk workers (9%)..

4.3.3 Summary of findings

Most of the demographics of our Indian Turk sample are similar to the U.S. Turk sample, except Indian MTurk workers reported higher levels of education. Almost everyone from the Indian Turk sample used social media. Indian MTurk workers reported having put more personal information online than the U.S. MTurk workers did. Although we might expect more use of social media and more information online to predict more privacy concerns (see Table 1 for the social media effect in the U.S. samples), this was not the case among Indian MTurk workers. They were less worried about their information and did not take more actions to protect their identity. Also, Indian MTurk workers showed less positive attitudes about anonymity than did U.S. MTurk workers. The only notable difference in the other direction is that Indian MTurk workers more often hid information from employers.

Indian MTurk workers' policy opinions were very different from those of U.S. MTurk workers. More than half thought their laws provide enough protection to their privacy, and more than half agreed to government monitoring. This difference might be due to cultural differences or a result of different national events or news. Additionally, there is a potential bias in that the surveys were conducted after the Snowden revelations (June 6, 2013 [3]). The news coverage of these revelations in the U.S. may have reduced American's trust in online privacy and government Internet policy and practices.

5. DISCUSSION

Consistent with previous research, our study shows that U.S. MTurk workers are younger than the general U.S. population and differ in other ways. But even controlling for demographic factors, more of these U.S. MTurk workers express worries and concerns about their online information. Moreover, U. S. MTurk workers are more likely to seek anonymity and be in favor of Internet policies that allow anonymity. Indian MTurk workers have weaker concerns about privacy.

One possible explanation for the differences between the U.S. MTurk and the U.S. public samples is that U.S. MTurk workers might be more technical savvy than the general public. We were unable to assess this possibility because the U.S. representative survey did not ask respondents about their knowledge of the Internet, or how much they used it. However, when asked about what information about them is online, the U.S. public sample showed more uncertainty than the U.S. Turk sample about what kinds of information is available about them online (mean pieces of information they are unsure about = 1.6 and 1.1, t [955] = 3.05,

$p < .01$), especially about their contact information (email addresses, phone numbers). This finding might indicate that the U.S. MTurk group has more privacy concerns about their personal information because they are more certain that others have potential access to it.

This work suggests that privacy researchers, in their studies using MTurk workers, may need to take into consideration the heightened privacy attitudes and behavior of the U.S. workers on MTurk. We provide quantitative evidence showing that U.S. MTurk workers often seek anonymity and have a heightened concern with privacy. Our results do not bear on the issue of internal validity of online experiments (e.g., [27]). Indeed prior work [8] suggests that internal validity of experiments using MTurk workers is similar to the validity of traditional lab experiments. What our results do suggest is that descriptive findings of privacy attitudes and behavior based on MTurk samples may not generalize to the broader population (i.e., external generalizability). Research (e.g., [29]) that uses crowds as a privacy evaluation platform should consider the potential sample bias when generalizing MTurk worker privacy preferences to other users.

We also found significant differences in opinions and experiences between MTurk workers recruited from the U.S. and India. Privacy researchers using MTurk should monitor and record the locations of their participants, and examine the effects of these differences.

6. CONCLUSION

The findings of our study suggest U.S. MTurk workers have similar amount of personal information online as the general American population, but they differ from the general public in their behaviors and opinions about online anonymity and privacy,. Indian MTurk workers have more personal information online than the U.S. MTurk workers, but have less preference towards anonymity and are less concerned about their privacy. Research on people's privacy opinions and preferences will need to account for differences between MTurk workers and the general public and perhaps introduce additional control variables to assess how extensive these differences are.

7. ACKNOWLEDGMENTS

Support for this work was provided by NSF grants CNS1040801 and CNS1221006. We thank the participants who helped pilot the survey.

8. REFERENCES

[1] "Millions of Target customers' credit, debit card accounts may be hit by data breach", http://www.nbcnews.com/business/consumer/millions-target-customers-credit-debit-card-accounts-may-be-hit-f2D11775203

[2] https://www.mturk.com/mturk/help?helpPage=policies

[3] http://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order

[4] Ackerman, M.S., Cranor, L.F., and Reagle, J. Privacy in e-commerce: Examining user scenarios and privacy preferences. *Proc. of the 1st ACM conference on Electronic commerce*, (1999), 1–8.

[5] Acquisti, A., Gross, R., & Stutzman, F. (2011). Faces of Facebook: privacy in the age of augmented reality. Presentation at BlackHat USA 2011, Las Vegas, NV, August 4th, 2011. http://www.heinz.cmu.edu/~acquisti/face-recognition-study-FAQ/

[6] Adjerid, A., Acquisti, A., Brandimarte, L., and Loewenstein, G. 2013. Sleights of privacy: framing, disclosures, and the limits of transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (SOUPS '13). ACM, New York, NY, USA.

[7] Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior research methods*,*43*(3), 800-813.

[8] Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. Political Analysis, 20(3), 351-368.

[9] Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. Perspectives on Psychological Science, 6(1), 3-5.

[10] Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. Computers in Human Behavior, 29(6), 2156-2160.

[11] Chen, G. M. (1995). Differences in Self-Disclosure Patterns among Americans Versus Chinese A Comparative Study. *Journal of Cross-Cultural Psychology*,*26*(1), 84-91.

[12] Christopherson, K.M. The positive and negative implications of anonymity in Internet social interactions. *Computers in Human Behavior 23*, 6(2007), 3038–3056.

[13] Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. Judgment and Decision Making,5(3), 159-163.

[14] Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. Journal of Behavioral Decision Making, 26(3), 213-224.

[15] Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples.Behavioral and Brain Sciences, 33(2-3), 94-95.

[16] Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. American Psychologist, 59(2), 93.

[17] Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. Identifying personal genomes by surname inference. *Science 339*, 6117 (2013), 321-324.

[18] Horton, J. J., & Chilton, L. B. (2010, June). The labor economics of paid crowdsourcing. In Proceedings of the 11th ACM conference on Electronic commerce (pp. 209-218). ACM.

[19] Ipeirotis, P. (2009). Turker demographics vs. Internet demographics. http://behind-the-enemy-lines.blogspot.-com/2009/03/turker-demographics-vs-Internet.html.

[20] Jensen, C. and Potts, C. Privacy practices of Internet users: self-reports versus observed behavior. *International Journal of Human-Computer Studies 63*, 1 (2005), 203–227

[21] Jensen, C. and Potts, C. Privacy policies as decision-making tools: an evaluation of online privacy notices. *Proc. of CHI 2004*, ACM (2004), 471-478.

[22] Kang, R., Brown, S., and Kiesler, S. 2013. Why do people seek anonymity on the Internet?: informing policy and design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). ACM, New York, NY, USA, 2657-2666.

[23] Kelley, P. G. (2010, July). Conducting usable privacy & security studies with amazon's mechanical turk. In Symposium on Usable Privacy and Security (SOUPS), Redmond, WA.

[24] Krishnamurthy, B. and Wills, C.E. Characterizing privacy in online social networks. *Proc. of the first workshop on Online social networks*, (2008), 37–42.

[25] Kumaraguru, P., & Cranor, L. (2006, January). Privacy in India: Attitudes and awareness. In *Privacy Enhancing Technologies* (pp. 243-258). Springer Berlin Heidelberg.

[26] Lasecki, W. S., Teevan, J., & Kamar, E. (2014). Information Extraction and Manipulation Threats in Crowd-Powered Systems. CSCW 2014.

[27] Lease, M., Hullman, J., Bigham, J. P., Bernstein, M., Kim, J., Lasecki, W., ... & Miller, R. C. (2013). Mechanical turk is not anonymous. Social Science Research Network. linklink

[28] Leon, P. G., Ur, B., Wang, Y., Sleeper, M., Balebako, R., Shay, R., ... & Cranor, L. F. (2013, July). What matters to users?: factors that affect users' willingness to share information with online advertisers. In Proceedings of the Ninth Symposium on Usable Privacy and Security (p. 7). ACM.

[29] Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., & Zhang, J. (2012, September). Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (pp. 501-510). ACM.

[30] Madden, M. and Smith, A. "Reputation management and social media." Pew Research Center's Internet & American Life Project. May 26, 2010. Available at:http://www.pewInternet.org/Reports/2010/Reputation-Management.aspx.

[31] Madden, M., Fox, S., Smith, A. "Digital footprints." Pew Research Center's Internet & American Life Project. December 16, 2007. Available at:http://www.pewInternet.org/Reports/2007/Digital-Footprints.aspx.

[32] Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). Being A Turker. Proc. of CSCW 2014.

[33] Mason, R. M., & Dupuis, M. J. (2014). Cultural Values, Information Sources, and Perceptions of Security. In iConference 2014 Proceedings (p. 778 - 783). doi:10.9776/14367

[34] Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. Behavior research methods, 44(1), 1-23.

[35] Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. Judgment and Decision making, 5(5), 411-419.

[36] Poole, E.S., Chetty, M., Grinter, R.E., and Edwards, W.K. More than meets the eye: transforming the user experience of home network management. *Proc. DIS 2008*, ACM Press (2008), 455–464.

[37] Pew Research Center (2013, June). Majority Views NSA Phone Tracking as Acceptable Anti-terror Tactic. http://www.people-press.org/2013/06/10/majority-views-nsa-phone-tracking-as-acceptable-anti-terror-tactic/

[38] Stuart, H.C., Dabbish, L., Kiesler, S., Kinnaird, P., and Kang, R. Social transparency in networked information exchange: a theoretical framework. *Proc. of CSCW 2012*, ACM (2012), 451–460.

[39] Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS One*, *6*(3), e16836.

[40] Rainie, L., Kiesler, S., Kang, R., Madden, M., Duggan, M., Brown, S., & Dabbish, L. (2013). Anonymity, Privacy, and Security Online. Pew Research Center.

[41] Rhee, H. S., Kim, C., & Ryu, Y. U. (2009). Self-efficacy in information security: Its influence on end users' information security practice behavior. *Computers & Security*, *28*(8), 816-826.

[42] Ross, J., Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10). ACM, New York, NY, USA, 2863-2872.

[43] Suler, J. The online disinhibition effect. Cyberpsychology & behavior 7, 3 (2004), 321–326.

[44] Teich, A., Frankel, M. S., Kling, R., & Lee, Y. C. (1999). Anonymous communication policies for the Internet: Results and recommendations of the AAAS conference. *The Information Society*, *15*(2), 71-77.

[45] Ur, B., & Wang, Y. (2013, May). A cross-cultural framework for protecting user privacy in online social media. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 755-762). International World Wide Web Conferences Steering Committee.

[46] Wang, Y., Norice, G., & Cranor, L. F. (2011). Who is concerned about what? A study of American, Chinese and Indian users' privacy concerns on social network sites. In Trust and trustworthy computing (pp. 146-153). Springer Berlin Heidelberg.

[47] Zhao, C., Hinds, P., & Gao, G. (2012, February). How and to whom people share: The role of culture in self-disclosure in online communities. In*Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 67-76). ACM.

APPENDIX

SURVEY QUESTIONS

Note: We only show the questions analyzed in this paper. Questions that were the same in the two surveys are numbered only (without any letters preceding the numbers). Questions that were different in the two surveys are marked using letters before the number (e.g., Pew survey items are designated "PEW", MTurk items are marked as "MTURK").

MTURK 1. Do you ever use a site like Twitter, Facebook, LinkedIn, Google Plus, or another social networking site? ☐Yes ☐No

PEW 1. Please tell me if you ever use the Internet to do any of the following things. Do you ever use the Internet to_____?

	Yes	No
Use a social networking site like Facebook, LinkedIn or Google Plus	☐	☐
Use Twitter	☐	☐

2. Is any of the following information about you available on the Internet for others to see? It doesn't matter if you put it there yourself or someone else did so.

	Yes, it's online	No, it's not online	Not sure	Does not apply
Your email address	☐	☐	☐	☐
Your home address	☐	☐	☐	☐
Your home phone number	☐	☐	☐	☐
Your cell phone number	☐	☐	☐	☐
Your employer or a company you work for	☐	☐	☐	☐
Your political party or political affiliation	☐	☐	☐	☐
Something you've written that has your name on it	☐	☐	☐	☐
A photo of you	☐	☐	☐	☐
Video of you	☐	☐	☐	☐
Which groups or organizations you belong to	☐	☐	☐	☐
Your birth date	☐	☐	☐	☐
Other information (please specify)				

3. Do you ever worry about how much information is available about you on the Internet, or is that not something you worry about? ☐Yes, worry about it. ☐No, don't worry about it. ☐Not sure

4. Considering everything you know and have heard about the Internet, do you think it is possible for someone to use the Internet completely anonymously – so that none of their online activities can be easily traced back to them? ☐Yes ☐No ☐Not sure

5. Have you ever tried to use the Internet in a way that hides or masks your identity from certain people or organizations? ☐Yes ☐No ☐Not sure

MTURK6. Do you ever post comments, questions, or information on the Internet using the following types of names?

	Yes	No	Not sure
Your real name	☐	☐	☐
A username or screenname that people associate with you	☐	☐	☐
A username or screen name that people do not associate with you	☐	☐	☐
No name at all	☐	☐	☐

PEW5. Do you ever post comments, questions, or information on the Internet _____?

	Yes	No
Using your real name	☐	☐
Using a username or screen name that people associate with you	☐	☐
Without revealing who you are	☐	☐

MTurk 7. Have you ever tried to use the Internet in such a way that your family members, a romantic partner, certain friends, coworkers would be unable to see what you have read, watched, or posted online? ☐Yes, I've done this. ☐No, I haven't done this.

MTurk 8. Have you ever tried to use the Internet in such a way that an employer, supervisor, or companies you work for would be unable to see what you have read, watched, or posted online? ☐Yes, I've done this. ☐No, I haven't done this.

MTurk 9. Have you ever tried to use the Internet in such a way that people from your past, or people who might criticize, harass, or target you would be unable to see what you have read, watched, or posted online? ☐Yes, I've done this. ☐No, I haven't done this.

MTurk 10. Have you ever tried to use the Internet in such a way that law enforcement, the government, or companies or people that might want payment for the files you download such as songs, movies, or games would be unable to see what you have read, watched, or posted online? ☐Yes, I've done this. ☐No, I haven't done this.

MTurk 11. Have you ever tried to use the Internet in such a way that hackers, criminals, or advertisers would be unable to see what you have read, watched, or posted online? ☐Yes, I've done this. ☐No, I haven't done this.

PEW 7. Have you ever tried to use the Internet in ways that keep _____ from being able to see what you have read, watched or posted online?

	Yes, did this	No, did not
Family members or a romantic partner	☐	☐
Certain friends	☐	☐
An employer, supervisor, or coworkers	☐	☐
The companies or people who run the website you visited	☐	☐
Hackers or criminals	☐	☐
Law enforcement	☐	☐
People who might criticize, harass, or target you	☐	☐
Companies or people that might want payment for the files you download such as songs, movies, or games	☐	☐
People from your past	☐	☐
Advertisers	☐	☐
The government	☐	☐

12. Thinking about current laws, do you think the laws provide reasonable protections of people's privacy about their online activities? ☐ Yes, they provide reasonable protection ☐ No, they're not good enough ☐ Not sure

13. Do you think that people should have the ability to use the Internet completely anonymously for certain kinds of online activities? ☐ Yes, should have the ability ☐ No, should not have the ability ☐ Not sure

MTurk 14. Do you think the government should be able to monitor everyone's email and other online activities if officials say this might prevent future terrorist attacks? ☐ Yes, should monitor ☐ No, should not monitor ☐ Not sure

---

These following questions are for statistical purposes only.

15. What is your gender? ☐ Male ☐ Female ☐ Other

16. How old are you (years)? _____

17. What is the highest level of school you have completed or the highest degree you have received?

☐ Less than high school (Grades 1-8 or no formal schooling)
☐ High school incomplete (Grades 9-11 or Grade 12 with NO diploma)
☐ High school graduate (Grade 12 with diploma or GED certificate)
☐ Some college, no degree (includes some community college)

☐ Two year associate degree from a college or university
☐ Four year college or university degree/Bachelor's degree (e.g., BS, BA, AB)
☐ Some postgraduate or professional schooling, no postgraduate degree
☐ Postgraduate or professional degree, including master's, doctorate, medical or law degree (e.g., MA, MS, PhD, MD, JD)
☐ Not sure

MTurk 18. Where were you born?

☐ China
☐ India
☐ United Kindom
☐ United States
☐ Other (please specify)_____

MTurk 19. Do you usually access the Internet from these locations?

	True	False	I'm not sure
China	☐	☐	☐
India	☐	☐	☐
United Kingdom	☐	☐	☐
United States	☐	☐	☐
Other (please specify)_____			

Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google

Emilee Rader
Department of Media and Information
College of Communication Arts and Sciences
Michigan State University
emilee@msu.edu

ABSTRACT

Internet companies record data about users as they surf the web, such as the links they have clicked on, search terms they have used, and how often they read all the way to the end of an online news article. This evidence of past behavior is aggregated both across websites and across individuals, allowing algorithms to make inferences about users' habits and personal characteristics. Do users recognize when their behaviors provision information that may be used in this way, and is this knowledge associated with concern about unwanted access to information about themselves they would prefer not to reveal? In this online experiment, the majority of a sample of web-savvy users was aware that Internet companies like Facebook and Google can collect data about their actions on these websites, such as what links they click on. However, this awareness was associated with lower likelihood of concern about unwanted access. Awareness of the potential consequences of data aggregation, such as Facebook or Google knowing what other websites one visits or one's political party affiliation, was associated with greater likelihood of reporting concern about unwanted access. This suggests that greater transparency about inferences enabled by data aggregation might help users associate seemingly innocuous actions like clicking on a link with what these actions say about them.

1. INTRODUCTION

In February 2012, the New York Times published an article describing how the Target Corporation uses "predictive analytics" to find patterns in personal information about customers and their behavior, that has been collected first-hand by Target or purchased from third parties [10]. The article continues to be frequently mentioned because of a (perhaps apocryphal) anecdote about a father who found out that his teenage daughter was pregnant, by looking through the coupons she received from Target via the US postal service. Over the past few years, this example has been used by many as a warning about the future of information privacy, because it illustrates how behavioral data that is collected without a person's knowledge as they interact with systems in their daily lives (here, purchase records from Target) can be used to infer intimate details that one might prefer not to disclose.

Most web pages include code that users cannot see, which collects data necessary for making predictive inferences about what each individual user might want to buy, read, or listen to[1]. This data ranges from information users explicitly contribute, such as profile information or "Likes" on Facebook, to behavioral traces like GPS location and the links users click on, to inferences based on this data such as gender and age [15], sexual orientation [18], and whether or not one is vulnerable to depression [7].

Whether or not users explicitly intended to provide the information, once it has been collected it is not just used to reflect users' own likes and interests back through targeted advertisements. Algorithms use this data to turn users' likenesses into endorsements—messages displayed to other users that associate names and faces with products and content they may not actually want to endorse [31, 32]. Algorithms make inferences about who we are, and present that information on our behalf to other people and organizations.

Internet users express discomfort with data collection that enables personalization. For example, a recent Pew survey found that "73% of search engine users say they would NOT BE OK [sic] with a search engine keeping track of searches and using that information to personalize future search results, because it is an invasion of privacy" [28]. Eighty-six percent of Internet users have taken some kind of action to be more anonymous when using the web—most often, clearing cookies and browser history [30].

Nevertheless, people use search engines and social media on a daily basis, and simple browser-based strategies like deleting cookies and browsing history are not enough to protect one's information online. For example, the configuration of plugins and add-ons of a particular web browser on a specific machine comprises a unique "fingerprint" that can be traced by web servers across the web, and this information is conveyed through headers that are automatically exchanged by every web browser and web server behind the scenes [25].

It is clear that users are concerned about online privacy, and that transparency—especially regarding what can be inferred about users based on seemingly innocuous data like clicking a link in a web page—is lacking. What, then, are the disclosures that users actually do know about, and how is this awareness related to privacy concern? The goal of this research was to investigate whether users recognize that their behaviors provision information which may be used by personalization and recommendation algorithms to infer things about them, and if this awareness is associated with privacy concern.

I found that a sample of web-savvy users were resoundingly aware that Internet companies like Facebook and Google can col-

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

[1] https://www.eff.org/deeplinks/2009/09/online-trackers-and-social-networks

lect data about their behaviors on those websites, consisting of things like when and how often they visit those sites, and what links they click on. I refer to information like these examples as *First Party Data*, because it can be collected directly from user actions with websites. However, greater awareness of the collection of First Party Data was associated with a LOWER likelihood of concern about unwanted access to private information.

Participants were much less aware of automatic collection of personal information produced by aggregation across websites, which can reveal patterns in what other websites such as one's purchase habits, or aggregation across users, which can reveal potentially sensitive information like sexual orientation. But unlike First Party Data, those users who had greater awareness of the either kind of aggregation had a GREATER likelihood of concern about unwanted access. This suggests that a solution involving informed consent about collection of First Party Data would not support better boundary management online, and that different approaches are needed to make the consequences of aggregation, rather than the disclosures themselves, more transparent.

2. RELATED WORK

2.1 Boundary Management Online

People interact with one another in contexts structured by the roles they assume and the activities they engage in; by the social norms of the situation; by their own objectives and goals; and even by aspects of the architecture of the physical world [26]. Westin [42] defined privacy as "the claim of an individual to determine what information about himself or herself should be known to others", and all of these factors contribute to people's assessments of what information they want to allow others to know in what context.

While there are many structural aspects of offline physical and social contexts that help people negotiate boundaries between public and private, managing boundaries when sharing information online is more difficult. Social media systems, in particular, suffer from "context collapse": users have multiple audiences for their posts with whom they might want to share different sets of information, but it can be difficult to understand which part of one's potential audience is able to see the content [12], or is even paying attention [29]. Stutzman and Hartzog [39] conducted an interview study of users with multiple social network profiles, who used profiles on different systems to manage boundaries and disclosures. They sometimes kept the profile identities completely separate, and other times they strategically or purposefully linked them to create boundaries between audiences with which they shared different degrees of intimacy. Different systems have implemented interface mechanisms and controls for specifying the boundaries between audiences, but no industry best practices or standards seem to exist for interfaces to manage access to one's personal information [4]. For example, Bonneau and Preibusch reported that at the time of their research, only two out of 45 social network sites (Facebook and LinkedIn) offered users the capability to see what their profile looked like to users with different levels of access.

Users don't always change privacy settings and mechanisms from the defaults, and even when they do, they aren't always successful at achieving their desired result. Liu et al. [21] designed a Facebook app to collect 10 photos from participants' Facebook accounts, along with the visibility setting associated with each photo. They also asked each user to indicate who their desired audience was for each photo. They found that 36% of the photos were shared with the default—fully public—setting, while participants indicated only 20% of the photos should have been public. In an experiment, Egelman et al. [11] presented users with different information sharing scenarios in Facebook and asked to specify access control polities. They found that when users made mistakes—when their desired level of access did not match what they specified through the system—they erred on the side of revealing more broadly than they wanted to.

In systems that do not provide privacy mechanisms, users express discomfort about what others might infer about them by learning about characteristics of the content they consume. Personalized content can reveal potentially embarrassing information to others [40]. For example, Silfverberg et al. [33] studied the social music service Last.fm and found that participants reported making personal judgments about other users based on their music preferences. Music has an emotional quality, and participants worried that allowing others to know what music they were listening to might reveal information about what they were feeling that they might not want to disclose. At that time, Last.fm did not allow users to protect any of the information in their profile, so the only recourse they had was to create separate profiles for different audiences.

Some users also express concern about the possibility that behavioral advertising might reveal private information about them based on past web browsing sessions. After having behavioral advertising explained to them, 41 out of 48 participants in one study felt concerned about what they perceived as a loss of control over their information [41]. A majority of participants in another study reported that they had been embarrassed in the past by advertising that appeared on a web page they were viewing, that was also seen by another person in the vicinity (e.g., "what were you browsing last night") [1]. These examples each illustrate circumstances where data collected for personalization has made it more difficult for users to manage the boundary between information they do and do not want to reveal.

2.2 Information vs. Social Privacy

There is an important distinction between *social privacy* and *information privacy*. "Social privacy" concerns how we manage self-disclosures, availability, and access to information about ourselves by other people. "Information privacy" refers to the control of access to personal information by organizations and institutions, and the technologies they employ to gather, analyze, and use that information for their own ends [36].

Privacy settings in most online systems are designed to manage social privacy, and people are willing to take steps to enforce social boundaries online when such options are available [16]. For example, people who are more concerned about information privacy reported using privacy management tools more, according to Litt [20] who analyzed a Pew Internet & American Life data set from 2010. However, people may not perceive a connection between social privacy and threats to information privacy. Strategies such as specifying one's privacy settings and maintaining multiple profiles allow users control over social privacy, but they do not support better control over information privacy, because the architectures and algorithms that collect and make inferences from the information are mostly invisible to users. It is difficult to manage information boundaries appropriately when users are unaware of disclosures [8].

While some of the information used by personalization algorithms for tailoring content to user interests and preferences comes from information people explicitly contribute and can therefore self-censor, much of the data is collected invisibly as users surf the web. Companies are not always as transparent as they could be in their stated practices about what data they have access to, and how they will use it. For example, Willis et al. [43] conducted an investi-

gation to determine the extent of personalization in Google search results. They "induced" interests in fake profiles by doing searches with particular keywords and viewing specific videos on YouTube, expecting that this information would be used by Google to determine which ads to display. Google's policy at the time stated that ads displayed with search results would be *contextual* ads, selected only based on information in the search result page itself. The researchers found that *non-contextual* ads based on inferred interests from previous interactions appeared alongside the contextual ads, despite the policy. They also found that some of the non-contextual ads could potentially reveal sensitive personal characteristics based on the inferred interests, such as an ad which contained the question, "Do you have diabetes?"

In a different study, Korolova [17] investigated the extent to which information Facebook users specified as available to "Only me" could be used for targeted advertising. In one example, she created a series of Facebook advertisements targeted toward characteristics of a person known to the research team, who had specified that profile information about age should be hidden from everyone. The specially crafted ads differed according to only one dimension: the age of the user to whom the ads should be displayed. Using Facebook's advertiser interface, Korolova was able to infer the private age of the target person based on updates about the performance of ad campaigns—since the ads for the incorrect ages were not displayed. Her experiment demonstrates the possibility that even when users indicate they want to keep specific information private, Facebook has used that information to target advertisements in a potentially revealing way.

In some studies, users report that they like personalized search, because personalization provides better results [27]. Likewise, many people say that they are comfortable with customized ads based on the contents of their email or Facebook profile, and also find tailored ads to be useful [1, 41]. However, when asked directly about the sensitivity of specific Google search queries, 84% of users in one study said that there were queries in their search history that they felt were "sensitive", and 92% wanted control over what Google was tracking about them as they searched the web [27]. Less than 30% of participants in another study were aware that browsing history and web searches could be used to automatically create a profile about them, and most people were unable to distinguish between the company represented by the ad content, and the company responsible for displaying the ad [41].

Altman [2] wrote, "If I can control what is me and not me; if I can define what is me and not me; if I can observe the limits and scope of my control, then I have taken major steps toward understanding and defining what I am." There are few options for users who want to manage multiple identities with respect to systems or companies, rather than self-presentation to other people, for the purpose of maintaining separate personalization experiences. The invisibility of the architectures and algorithms responsible for personalization make it difficult for users to manage boundaries appropriately with respect to information privacy [8].

2.3 Research Questions

Users may be in danger of losing control over the mechanisms by which they develop and enforce their individuality online, because they don't know and can't control who the system thinks they are, and how that identity is presented to other people and organizations. This study focused on situations people encounter in everyday web use where information disclosure boundaries are not straightforward. The purpose was to investigate (1) whether users are concerned about privacy when they engage in common behaviors on the web that can enable automated disclosures to take place; (2) whether people are aware of different types of data that can be automatically collected about them when they use Facebook and Google Search; and (3) how the perceived likelihood of automated data collection might be related to privacy concern.

3. METHOD

I conducted a 2 (*Site*: Facebook or Google Search) x 3 (*Behavior*: Link, Autocomplete or Ad) x 2 (*Sensitivity*: High or Low) between-subjects online experiment hosted by Qualtrics, in May 2013. Participants viewed a hypothetical situation that varied according to these three dimensions, which are described in detail below. This study was approved as minimal risk by our Institutional Review Board.

3.1 The Site Dimension

The two levels of the *Site* dimension were Facebook and Google Search. Interacting via social media and searching for information on the web are two very common Internet-related activities, yet they have some interesting similarities and differences. Many of the underlying web technologies, particularly related to the implementation of dynamic, interactive web pages, are the same in these two situations. However, one way in which these two sites differ is the degree to which user actions take place in a social context. Searching is typically a solitary activity, and it is reasonable to assume that people feel more like they are interacting with the search engine database than another human being when they search for something. Using social media feels like communicating, even when one is simply browsing the Facebook News Feed. This contextual difference could affect whether people feel their actions on the two sites can be observed or not. In addition, the settings and mechanisms users have to control access to their information on Facebook are all geared toward social privacy, not information privacy.

3.2 The Behavior Dimension

I chose three behaviors to include in this study: clicking a link, typing in a text box, and viewing ads in a web page. These behaviors seem on the surface like they are not directly related to disclosures of personal information, because they do not directly ask for it. However, it is possible to infer personal information from all three.

Clicking a Link: When a user clicks a link in Facebook or Google, he or she sees visual feedback that the system has registered the action when the web page changes to display new content. Clicking a link in both systems sends a request to the server that hosts the content of the page the user is navigating to. Users may already be aware of this, since it is a fundamental aspect of how the Internet works. However, both Google and Facebook can employ redirects so that they can collect data about which links users click on. So while there is visible feedback that something server-related is happening, it is less clear to users that Google and Facebook can record information about what links you click on.

Data consisting of which links users have clicked on can be used to infer the gender and age of individual users who have not revealed that information, as long as a sufficient number of other users with similar browsing patterns have provided their gender and age information. This is accomplished by first identifying the most common gender and age segment for the visitors of a set of web pages. Then, the age and gender of other visitors to those pages are inferred, whether or not they have chosen to reveal them. Gender can be inferred with 80% accuracy, and age with 60% accuracy [15].

Typing and Autocomplete: When a user types in a text box on

Facebook or Google Search, both sites send individual characters back to the server as they are typed. This real-time communication supports auto-completing search terms and the names of Facebook friends when creating a status update, without having to explicitly click the Submit button. However, the extent to which this feedback might be understood to communicate outside the web browser differs across the two sites. For example, when a user types a status update, the only visual indicator that information has been transmitted occurs when one's Facebook friends' names appear below the text box. However, Google Instant Search updates the entire web page as a search query is typed by the user. These different levels of feedback may lead to different conclusions on the part of the user about what and how much information might be going back-and-forth between themselves and the system as they are typing, before they explicitly submit the text. In reality, data is sent back to the server in both cases.

Viewing Ads in a Web Page: Ads in web pages can have a visible relationship with other information displayed at the same time in the web page (called *contextual* ads), or be based on other data available to advertising companies about the end user (confusingly called *non-contextual* ads) [43]. Therefore, different types of ads provide different kinds of feedback from the system to the user about inferences the system has made about them. Google ads in search result pages appear after the user has requested information via a search query, and tend to be contextual. This might trigger users to notice that ads are personalized, and they might therefore be more concerned about privacy. On the other hand, because Facebook ads are more likely to be based on one's profile information and "Likes" rather than information displayed in the News Feed (i.e. non-contextual), users who notice this may feel more concern about why particular ads were selected for display. However, there is invisible data collected too, that users do not receive feedback about: when an ad loads in a particular page, data is recorded about which ad loaded where.

3.3 The Sensitivity Dimension

The sensitivity of the information involved might increase overall privacy concern, and affect whether users wonder if data about their actions can be recorded. The High Sensitivity condition included ads, links to content, and search queries or posts about depression, a psychological disorder that is both common and highly stigmatized, and affects both men and women [23, 13]. The content and statements in the stimulus materials related to depression were based on research conducted by Moreno et al. [24], looking at college students' references to their own depression on social media websites. The Low Sensitivity condition consisted of content such as links to the website of the a local minor league baseball team, a technology-related article, and ads for a laptop or iPad.

3.4 The Experiment Procedure

The online experiment started by displaying a hypothetical situation that varied by condition, designed to closely resemble common experiences while using the web. Below is the text displayed to participants, corresponding with the levels of the *Behavior* dimension. Each condition was accompanied by a partial screen capture to illustrate what was happening, and the manipulation of *Site* and *Sensitivity* took place via the screen captures. All screen captures are included in Appendix A.

Link You visit Facebook and start reading posts in your Facebook News Feed. You scroll down the page, and click on a link a Facebook Friend has shared. The page changes to show the web page for the link that you clicked on.

Autocomplete You visit Google and start typing in the search box. Google makes a guess about what you might be searching for, and shows search results before you finish typing.

Ad You are viewing posts in your Facebook News Feed. As you scroll down the page, reading posts made by Facebook friends, you notice ads displayed on the right side of the screen.

Participants were asked a closed-ended and an open-ended *privacy concern* question, immediately after viewing the hypothetical situation:

1. Would you be concerned about unwanted access to private information about you in this scenario? [Yes, Maybe, No]
2. Please explain your answer to the previous question. [open-ended]

This emphasis on "unwanted access" follows from several definitions of privacy as control over access [42, 2]. Asking participants about concern over unwanted access is essentially operationalizing privacy as control over one's information. Likert scales often measure both direction and intensity at the same time (e.g., a "Very Satisfied" to "Very Dissatisfied" scale measures both whether someone was satisfied or dissatisfied, and by how much) [9]; however, the privacy concern question in this study asks about the presence or absence of concern, not how much concern. The additional *Maybe* option, rather than simply *Yes* or *No*, allows more accurate measurement of responses by not forcing participants to choose between the two extremes if they were unsure.

Asking the question in this way does not ask participants about specific things that may have caused them concern, and therefore it is not clear what they might have been thinking about when they answered the question. This phrasing of the question was intentional, in order to avoid "priming" participants to consider things they might not have thought about before when answering the question. The point of the manipulation was to trigger participants to think about a specific situation, but NOT to trigger them to think about specific *characteristics* of the situation, as a way to get as unbiased a response at possible given the study format.

After the privacy concern question, participants responded to a 16-item question that asked them to estimate the likelihood that Facebook or Google could collect different kinds of data about them: "How likely do you think it is that [Google | Facebook] can AUTOMATICALLY record each of the following types of information about you?" The motivation for asking about these items was to identify what kinds of "tracking" users think may be going on when they use the web, and through later regression analysis to identify associations between these beliefs and the likelihood of privacy concern. Participants indicated the likelihood of each statement between 0 and 100 in intervals of 10, using a visual analog scale represented as a slider. Half of the participants in the study were asked these questions about Facebook, and the other half about Google, and this depended on what *Site* condition they were randomly assigned to after they completed the consent form. The 16 items ranged from the clearly possible (which links the user clicks on), to the unlikely to be perceived as possible to collect (what the user's desktop image looks like). The question also included a few examples of information that can be inferred; for example, sexual orientation, which can be inferred from Facebook "Likes" [18]. However, few participants were expected to believe it likely that Facebook or Google could automatically detect this. See Figure 6 for the text of the items.

I included two sets of *control questions* in the survey: one to measure participants' Internet literacy (operationalized as familiarity with a set of Internet-related terms), and another to gauge the level of importance each participant placed on digital privacy. The questions that comprise the Internet Literacy index variable are based on the Web Use Skills survey reported in Hargittai and Hsieh

	Ad		Autocomplete		Link	
	High	*Low*	*High*	*Low*	*High*	*Low*
Facebook	60	60	61	56	60	60
Google	59	55	61	55	60	54

Figure 1: Number of participants in each condition. Independent variables are Site (Facebook or Google), Behavior (Ad, Autocomplete, or Link), and Sensitivity (High or Low).

(2011) [14]). This variable consists of the average of participants' assessments of their level of familiarity with the a list of Internet-related terms (*M*=3.57; *SD*=0.75, Cronbach's α=0.8).

I selected the questions that make up the Privacy Preferences index variable from two published privacy scales. The first was the "Blogging Privacy Management Measure", an operationalization of Communication Privacy Management theory applied to blogging by college students by Child et al [5]. This scale measures how bloggers think about boundaries between private and public when disclosing information online. I modified 8 items from that scale, replacing "blog" with "Facebook" where appropriate. An example item included in this study is, "If I think that information I posted to Facebook really looks too private, I might delete it." In addition, I selected four items from the "Information Privacy Instrument" developed by Smith et al [37]. This scale was designed to measure individuals' perceptions of organizational practices surrounding information privacy. An example item from this scale used in the study is, "It usually bothers me when companies ask me for personal information." Participants responded to these 12 items on a 5-point likert scale of Strongly Disagree—Strongly Agree.

To create the index variable, I reverse-coded where necessary and averaged across all 12 questions. The Privacy Preferences index variable therefore represents both attitudes toward individual disclosure in social media, and comfort level with the way organizations handle private user data. The mean of the privacy preferences variable was 4.003 (*SD*=0.5, Cronbach's α=0.74), which indicates that on average, participants valued online privacy, and were bothered by the idea of companies selling information about them to third parties.

3.5 Participants

I recruited participants from Amazon Mechanical Turk (MTurk), and restricted the sample to workers from the USA who had a 95% or higher approval rating after completing at least 500 tasks. MTurk workers were first required to answer an eligibility screening questionnaire. Participation was limited to MTurk workers who reported that they visited both Facebook and Google Search at least weekly, and were 18 or older. Using web-savvy MTurk workers as participants was convenient, but also purposeful: people who make money by completing tasks on the Internet are a best-case scenario for finding a population that is aware of invisible data collection and privacy risks on the Internet, compared with the usual suspects like undergraduates or a snowball sample. Participants completed the questions in an average of 7.56 minutes (*SD*=6.1 minutes) and received $2 in compensation. 748 participants started the survey; 47 were excluded because they did not finish the survey, or they failed to answer the attention check questions correctly, or they completed the survey during a Qualtrics service disruption.

After these exclusions, the number of participants remaining in each condition ranged from 54 to 61 (see Figure 1). The answers of the remaining 701 participants to the demographic questions resemble what other researchers have found about MTurk samples [3]—this sample was young (*M*=30.25 years old, *SD*=9.22), 80% white, more male (57%) than female (42%), and the majority (79%) had completed some post-high-school education or earned a 4-year college degree. Nearly all participants reported visiting Facebook (86%) and Google Search (98%) daily or more often. Finally, 97% of participants in the final sample reported having personally experienced a situation similar to the condition they were assigned to in the study.

	Estimate	*Odds Ratio*	*Std. Error*
Behavior: Autocomplete	-1.86***	0.16	0.37
Behavior: Link	-1.03**	0.36	0.35
Site: Google	-0.80***	0.45	0.35
Sensitivity: Low	-0.28	0.75	0.35
Autocomplete x Google	1.28*	3.59	0.51
Link x Google	1.03*	2.80	0.49
Autocomplete x Low	-0.01	0.99	0.54
Link x Low	-0.24	0.79	0.50
Google x Low	-0.80	0.45	0.51
Autocomplete x Google x Low	0.22	1.24	0.76
Link x Google x Low	-0.48	0.62	0.75
Internet Literacy	-0.12	0.89	0.10
Privacy Prefs	0.99***	2.71	0.17

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1: Coefficients for the Proportional Odds Mulitnomial Logistic Regression. The dependent variable represents participants' level of concern over unwanted access to private information, with three levels: *Yes, Maybe,* and *No*. The Baseline condition is Facebook:Ad:High. AIC is 1309.42; McFadden's Pseudo-R^2 is 0.096.

4. RESULTS

As expected based on previous research, more people answered *No* (377 participants) and *Maybe* (173 participants) than *Yes* (151 participants) when asked if they were concerned about unwanted access to private information. What follows are several analyses that help us to better understand when participants were more likely to express concern.

4.1 Conditions and Privacy Concern

I used a Proportional Odds Multinomial Logistic Regression to evaluate the relationship between the experiment conditions (*Site* x *Behavior* x *Sensitivity*), Internet Literacy and Privacy Preferences as controls, and the dependent variable: participants' answers to a single question about whether they would feel concerned about unwanted access to private information in the condition they were randomly assigned to. Like any closed ended question having an ordinal response format, it is possible that a *Yes* from one participant might mean more concern than another participant's *Yes*. While it is impossible to objectively compare the subjective experience of concern across participants, within each individual it is reasonable to interpret *Yes* as more concern than *Maybe*, which is more concern than *No*. The results from the model are in Table 1.

The multinomial logistic regression estimates the probabilities of choosing higher levels of concern than *No*. The baseline condition is Facebook:Ad:High, and all of the coefficients must be interpreted in relation to that combination of categories. Positive coefficients indicate greater likelihood of expressing concern; coefficients around 0 mean no additional likelihood on top of the baseline, and negative coefficients indicate lower likelihood of concern. For example, the large, negative estimate for the Autocom-

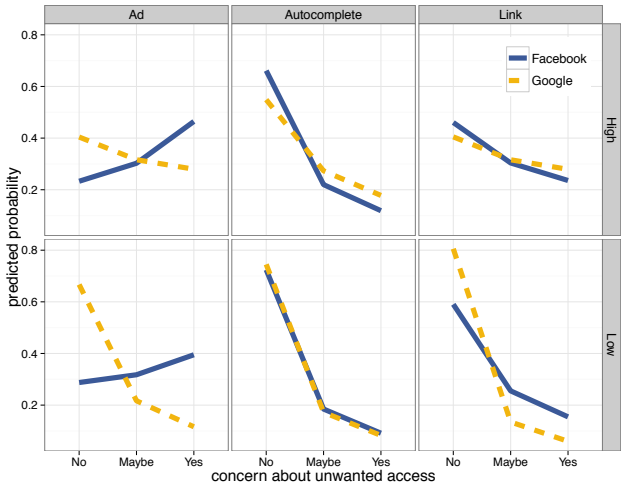


Figure 2: Predicted probabilities from the regression model presented in Table 1. The x-axis is the categorical response to the concern question, and the y-axis is the predicted probability of choosing a particular response.

plete conditions (-1.86) means that participants exposed to these conditions were much LESS likely to say they would be concerned about unwanted access to private information than participants exposed to any of the Ad conditions. Figure 2 presents the results as predicted probabilities generated from the model for a hypothetical participant who is average on the Internet Literacy and Privacy Preferences control variables.

Privacy Concern is Highest for Facebook Ads

Participants were most likely to express concern about unwanted access when they viewed the Facebook Ad conditions at both levels of Sensitivity. Participants who answered *Yes* to the concern question in the Facebook:Ad:High Sensitivity condition explained why they were concerned, by suggesting that the content of the ads makes them feel uncomfortable about what Facebook knows about them. They said things like, "Private information is being read from my posts," and "These ads seem to tell me that the computer knows about certain traits of mine due to my computer's history. I don't want Facebook to have this access." Participants in the Google:Ad:High Sensitivity condition expressed similar concerns, although fewer answered *Yes* to the concern question: "I would be concerned that someone could find out my search for depression by checking my Google search history, and that they keep a record of that when they display ads to me."

In contrast, participants in the Google:Ad:Low Sensitivity condition who said they would NOT be concerned about unwanted access said things like the following: "I think I've gotten used to having google [sic] searches causing ads to be pushed at me. In this case, nothing in the results is based on personal information–it's all from the search query just entered." This statement clearly expresses that the participant believes search results and ads are based on search queries, not personal information, implying that the participant feels the queries themselves are not personal information.

Figure 2 also clearly illustrates a statistically significant Scenario x Site interaction. Participants were more likely to say they were unconcerned than concerned about unwanted access to private information in the Google:Ad conditions. However, the opposite was true for participants exposed to the Facebook:Ad conditions. This means that web-savvy users, like Turkers, are more worried about privacy violations when they see targeted ads in Facebook than in Google Search.

Privacy Concern is Similar for Sensitive Ads and Links

The lines on the graph in Figure 2 for both Facebook and Google in the Link:High sensitivity conditions are similar to each other, and they also look very similar to the line for Google in the Ad:High condition. These predicted probabilities were indeed very similar: around 40-45% likelihood of answering *No*, 30-32% likelihood of answering *Maybe*, and 24-28% likelihood of answering *Yes*. In other words, participants were similarly likely to express concern about clicking on a "sensitive" link about depression in Facebook OR Google, as about viewing "sensitive" ads about depression in Google. Reasons they expressed for being concerned included statements focused on social, not information privacy: "Because, I just clicked on the link. I only would be concern if facebook [sic] announced on the news feed that I read the article"; and "it wouldn't bother me in the least if it was discovered that i'd [sic] been searching for information on depression". However, participants who did express concern said things that indicated they are aware of some of the data collected about them, e.g.: "I am very concerned about my search history, and specifically in this scenario I would be concerned about someone knowing I was depressed" and "Sometimes you get to stories by linking from other places online, and those could turn up in the URL of the story. Someone clicking on it could potentially figure out where I was surfing."

Privacy Concern is Lowest for Links in Google

The lowest likelihood of concern about unwanted access to private information in the experiment came from participants exposed to the Google:Link:Low Sensitivity condition. Just 6% of participants having average Internet Literacy and Privacy Preferences exposed to this condition are predicted by the model to choose *Yes*. This is clear evidence that web-savvy users view clicking on links in Google search results as an activity that does not have the potential to reveal information about them. As one participant explained, "It's just a link to a page. It's not asking for any personal information."

Autocomplete Does Not Warrant Concern

Participants in the Autocomplete conditions consistently reported that they would not be concerned about unwanted access to private information. Just 29 out of 233 participants exposed to Autocomplete conditions, across all levels of Site and Sensitivity, expressed concern. Their explanations made vague allusions to being tracked online, without being specific or technically accurate: "Nothing is every [sic] really private when online and Facebook offering suggestions when I type a status update proves I'm not just being paranoid."

The 155 participants in Autocomplete conditions who answered *No* to the privacy concern question gave reasons based on the *Site* they were asked about. Facebook participants in the Autocomplete condition who were unconcerned gave reasons such as, "I am not concerned about my privacy because Facebook already has my friends [sic] information. Facebook is just taking the list of my friends and presenting them in a new way." Likewise, participants exposed to both Google Autocomplete conditions said things like, "I don't really find this to be an invasion of privacy, I see it as Google thinking ahead. I would be pleased if the search that I wanted popped up before I finished typing it. It would save me some time"; and "The information that they are presenting is [the] most common used search that involves what you are beginning to

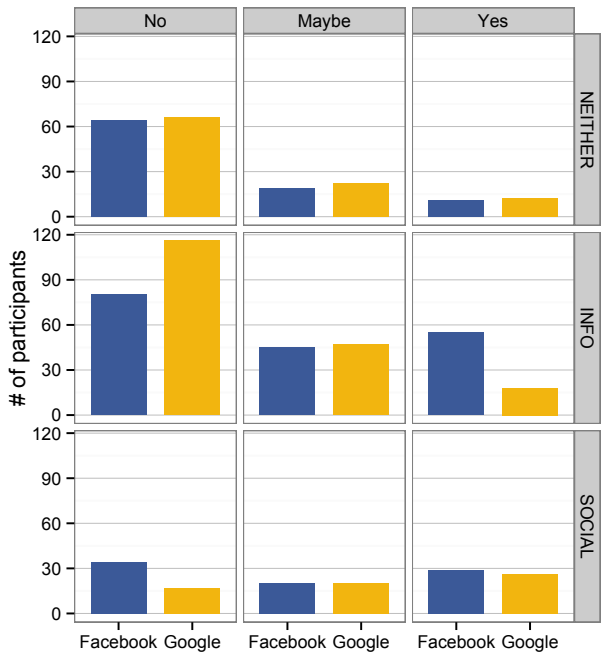


Figure 3: Number of responses coded as *Neither, Info* or *Social*, broken down by *Site* and the participant's concern response.

type. It does not contain specific information about what I have searched for."

In fact, Autocomplete works by sending keystrokes back to the servers of Facebook and Google, as they are typed, and matching them with other users' previously recorded queries. It is possible to use freely available "developer tools" for popular web browsers (e.g., Firebug, a plugin for Firefox) to see requests that pass information back and forth between the browser and Facebook's or Google's servers. On Facebook, this includes each character as it is typed in the Status box. These requests happen in the background, very quickly, and are typically not visible to end users. Features like Autocomplete further blur the line between social vs. information privacy, and recent research about self-censorship in social media [6, 35] does not take into consideration that users share ALL content they type with Facebook and Google, not just what they choose to submit or post.

"Unwanted Access" Refers to Websites, Companies

It is possible that when two different people answered *Yes* to being concerned about unwanted access to private information, they were concerned about different things. To investigate this, I analyzed participants' open-ended explanations for why they chose *Yes, Maybe* or *No* to the privacy concern question, to better understand what participants interpreted "unwanted access" to mean. A research assistant who had not previously examined data from this study used a bottom-up process to identify themes in 100 randomly selected responses, and developed the coding scheme based on those themes. The research assistant and the author then coded all 701 responses, without knowing which condition each response had come from or how the participant had answered the privacy concern question. The coders met to resolve disagreements and produce a final coding for each response. Cohen's κ was 0.82, indicating "excellent" inter-rater agreement [19].

	Estimate	*Odds Ratio*	*Std. Error*
Site: Google	0.116	1.123	0.306
Code: INFO	1.043***	2.839	0.264
Code: SOCIAL	1.136***	3.115	0.305
Google x INFO	-1.135**	0.321	0.371
Google x SOCIAL	0.374	1.454	0.437
Internet Literacy	-0.059	0.942	0.101
Privacy Prefs	0.922***	2.515	0.165

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2: Coefficients for the Proportional Odds Mulitnomial Logistic Regression. The dependent variable represents participants' level of concern over unwanted access to private information, with three levels: *Yes, Maybe,* and *No*. The Baseline condition is Facebook:NEITHER. AIC is 1334.33; McFadden's Pseudo-R^2 is 0.070.

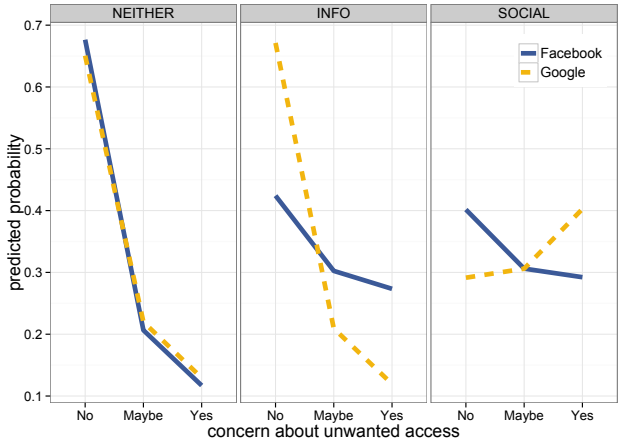


Figure 4: Predicted probabilities for the regression in Table 2. The x-axis is the categorical response to the concern question, and the y-axis is the predicted probability of choosing a particular response.

The final coding scheme had three mutually-exclusive categories, *Neither, Info* or *Social*. Responses coded as *Neither* did not provide enough evidence for coders to tell what kind of access the participant focused on when deciding whether he or she would feel concerned in the hypothetical situation. Examples of responses coded as *Neither* (n=194) include, "Nothing on the Internet is really private" and "All that appears is my name and where I am".

Responses coded as *Social* (n=146, the smallest category) included language referencing control over access by specific people, such as friends and family, social network connections, work supervisors, or being targeted by hackers. Responses coded *Social* were similar to the following: "No reason to be afraid, especially if my friend wouldn't mind it" or "I hate when previous searches pop up while someone is browsing my computer."

Finally, responses coded as *Info* (n=361, the largest category) mentioned control over access by websites, companies, governments, or other organizations. More responses were coded *Info* than *Social* or *Neither* combined. Many of these responses used passive voice and ambiguous pronouns, indicating that it may have been difficult for participants to put into words specifically when or how the unwanted access could take place. Examples of *Info* responses include, "I wouldn't really be offended by them targeting

ads towards me. That's how they make money" and "I wouldn't be 100% sure that my information was not linked to this site when I clicked the link."

In a few instances, responses contained both references to information and social privacy. If it was possible to tell which type of unwanted access the participant was more concerned about, that code was applied; otherwise, these responses were coded as *Social* (this happened only a handful of times). The number of responses coded as each category is presented in Figure 3, broken down by *Site* and the participant's concern response.

More "Info" Concern about Facebook than Google

I conducted a Proportional Odds Multinomial Logistic Regression with concern about unwanted access as the dependent variable, *Site* and *Type of Unwanted Access* (Info or Social) as regressors, and Internet Literacy and Privacy Preferences as controls. This analysis allows me to estimate, for example, the likelihood that a participant who mentioned social versus information privacy in his or her explanation would report concern about unwanted access depending on exposure to hypothetical situations involving Facebook or Google. The regression results are presented in Table 2.

The large, positive coefficients for the *Info* and *Social* categories mean that responses assigned those codes were more likely to be associated with *Yes* answers to the concern question, than responses coded as *Neither*. The large, negative coefficient for the Google x *Info* category means that information privacy concern was less likely to be associated with *Yes* answers in the Google conditions than in the Facebook conditions. All of these coefficients are also statistically significant.

The graph in Figure 4 shows the predicted probability of concern for participants with average Internet Literacy and Privacy Preferences. This graph illustrates that when participants associated "unwanted access" with privacy from websites, companies, and other institutions, those who were randomly assigned to Facebook conditions (solid blue lines in the graph) were more likely to express concern than those assigned to Google conditions (yellow dotted lines). However, this pattern was reversed for participants that associated "unwanted access" with social privacy. Participants who mentioned privacy from other people in the explanations for their answers were more likely to say they would be concerned when exposed to hypothetical situations involving Google than Facebook.

4.2 Perceived Likelihood of Data Collection

I conducted an exploratory factor analysis to identify patterns in participants' perceived likelihood that different types of data can be collected about them automatically while interacting with Facebook or Google Search. The maximum likelihood extraction with varimax rotation resulted in four interpretable factors. The factor loadings and text of the items are in Figure 6, and frequency histograms for each item are represented in Figure 5. The x-axis of each histogram in Figure 5 represents participants' assessments of the likelihood of each type of data being collected about them, ranging from 0 (Unlikely) to 100 (Likely) in increments of 10. The y-axis represents the number of subjects who chose each likelihood increment, for each variable. The gray line represents Facebook, the black dotted line in each histogram represents Google. Reliability scores (Cronbach's α) are also reported in Figure 6, for index variables created for each factor by averaging within participants across all items that comprised the factor.

OLS regressions with each factor's index variable as the dependent variable and the experiment conditions plus Internet Literacy and Privacy Preferences as controls revealed no significant interactions. This means that participants' answers on these items did not vary based on the experiment condition they were randomly assigned to. However, there was a main effect for *Site*, likely because participants were asked to estimate the likelihood of automatic data collection in Facebook OR Google. (Participants assigned to one of the Google conditions answered questions about Google throughout the entire study.)

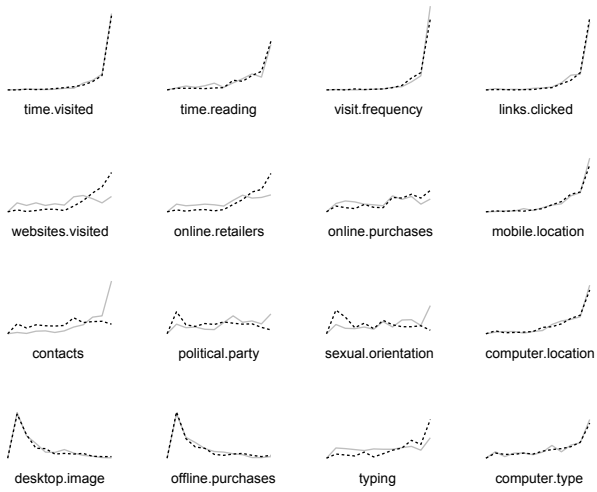


Figure 5: The x-axis of each frequency histogram represents participants' judgments of the likelihood of each type of data being collected about them, ranging from 0 (Unlikely) to 100 (Likely). The y-axis represents the number of subjects who chose each likelihood increment. The gray lines represent Facebook, the black dotted lines, Google. The questions associated with each histogram are in Figure 6.

Factor 1: First-Party Data

The questions that make up the "First-Party Data" factor are across the top of Figure 5 and down the right side. This factor includes the items *time.visited, time.reading, visit.frequency, links.clicked, mobile.location, computer.location* and *computer.type*. Each item asks about information that is available to websites directly as a result of user interaction. The pattern of these responses clearly illustrates that participants were aware that these types of information can be automatically collected. Nearly every participant felt that what time they visited Facebook or Google could be collected, for example, but there was a little bit more variance among participants about whether it is likely that Facebook or Google could figure out what type of computer they were using. It is actually possible to automatically collect this information—one's operating system and browser version are sent from the web browser to the web server when it requests a page.

Factor 2: Aggregation Across Sources

The questions making up Factor 2, "Aggregation Across Sources", are displayed in the first three histograms of the second row of Figure 5. Items *websites.visited, online.retailers* and *online.purchases* represent information about what other websites one visits and what kinds of things one shops for online. This is information Facebook and Google can only know by partnering with other websites, and associating one's profile with his or her behavior on those sites. This kind of data is similar to what one might see in a credit report that aggregates financial activity across multiple accounts, but without the score, and realize that it is possible to obtain a history

	Alpha	*Factor Loading*	*Abbreviation*	*Mean*	(*SD*)
First-Party Data	**0.78**			**84.9**	(**14.2**)
what time of day you visit [Google | Facebook]		0.817	*time.visited*	92.0	(15.6)
your physical location when using [Google | Facebook] on a mobile device		0.506	*mobile.location*	84.9	(19.9)
how much time you spend reading [Google | Facebook]		0.526	*time.reading*	80.0	(25.5)
what kind of computer you are using when you visit [Google | Facebook]		0.412	*computer.type*	71.8	(30.6)
your physical location when using [Google | Facebook] on a computer		0.501	*computer.location*	81.2	(23.9)
how often you visit [Google | Facebook]		0.756	*visit.freq*	93.2	(13.9)
what links you click on in your [Google search results | Facebook news feed]		0.712	*links.click*	91.0	(16.2)
Aggregation Across Sources	**0.87**			**67.0**	(**22.7**)
what websites you visit most often		0.764	*websites.visited*	69.6	(29.8)
which online retailers (e.g. Amazon.com) you visit most often		0.931	*online.retailers*	71.1	(29.0)
what you purchase from online shopping websites		0.689	*online.purchases*	60.1	(31.2)
Aggregation Across People	**0.80**			**57.0**	(**27.7**)
which people you communicate with online most often		0.548	*contacts*	70.0	(30.5)
your political party affiliation		0.815	*political.party*	50.8	(32.7)
your sexual orientation		0.860	*sexual.orientation*	51.0	(34.7)
"Impossible" to Collect	**0.60**			**19.4**	(**20.8**)
what the desktop image on your computer looks like		0.651	*desktop.image*	19.0	(24.0)
what you purchase from a brick-and-mortar store		0.477	*offline.purchases*	19.7	(25.1)
Not part of any factor					
what you are typing in the [search | Post or Comment] box before you submit		*n/a*	*typing*	65.0	(32.9)

Figure 6: Items measuring participants' beliefs about the likelihood that different types of data can be collected about them automatically by Facebook or Google [0 (Unlikely) to 100 (Likely)]. These items were presented in random order to each participant; here they are grouped and labeled according to the results of an exploratory factor analysis. Cronbach's α reliability scores are presented for each factor.

of one's activity that would be difficult to reconstruct from memory.

Participants were more divided in their judgments about the likelihood that Facebook and Google can know things about them that require this kind of aggregation. Participants assigned to Google thought it was more likely that information about what websites they visit and where they shop online could be collected, than participants assigned to Facebook. Interestingly, the technology and business partnerships with data aggregators that are necessary to collect this kind of data are feasible and practiced by practically all websites that use advertising. The variability in these responses indicates that participants estimations of likelihood are not likely to be based on knowledge about what is technically possible.

Factor 3: Aggregation Across People

Participants asked about Facebook vs. Google diverged the most on the items that make up the "Aggregation Across People" factor. The histograms for these questions are represented in the third row of Figure 5. This factor consists of one's *contacts, political.party,* and *sexual.orientation*: information that can be inferred through comparing patterns of behavior across people. For example, if some people disclose their sexual orientation directly in their profile, others with similar behavior patterns that did not choose to reveal this information may still be labeled the same. This kind of data is like the score or rating part of one's credit report, in that it provides information about how the system evaluates one's activity in the context of other people.

Participants asked about Google were spread across the range of responses for these questions, but tended toward thinking that it was unlikely Google could automatically collect information about their political party affiliation or sexual orientation, or the people they communicate with online. Participants who answered the questions about Facebook reported higher estimates of likelihood that this information could be automatically collected. All three of these types of information can actually be inferred from information users disclose online.

Factor 4: "Impossible" to Collect

Factor 4 consists of only two questions, that stand out in the bottom left corner of Figure 5 as the only two questions that skew toward the left or "unlikely" end of the range of possible responses, indicating that most participants believed it is not likely that Facebook or Google can automatically collect this information. This factor includes questions about the desktop image on one's computer and purchases in brick-and-mortar stores (*desktop.image, offline.purchases*). In fact, through partnerships with data aggregators it is possible that web companies can access data about users' buying habits in brick-and-mortar stores [34]. However, while it is technically possible for a web company to detect what a computer's desktop image looks like, it would be difficult to accomplish without compromising the security of the computer. I included the *desktop.image* question as a way to anchor the interpretation of users' responses to the awareness questions; if many participants thought this was possible, all responses to questions in this section of the survey would be suspect.

Typing

Finally, one question was not part of any factor: the likelihood that Google and Facebook can automatically collect "what you are typing in the [search | Post or Comment] box before you submit". Participants who answered questions about Facebook were fairly evenly spread across the range of responses (M=55.24, SD=33.7), indicating that participants varied in their beliefs about whether Facebook can record users' keystrokes as they are typing. However, the pattern is different for Google: more participants who answered the version of the question about whether Google can automatically collect information about what they are typing before

they submit the text reported feeling that this data collection was likely (*M*=75.17, *SD*=28.66).

Responses to this question are an indication that the nature of the interaction, and the type of visual feedback, may be important for understanding what is going on "under the hood". Google Instant Search provides search results as users type, and the entire web page updates to reflect search results. This seems to convey to at least some web-savvy users that information they are typing is been sent to Google in real-time. However, the information Facebook displays as users are typing consists of the names of one's friends that match the characters that have been typed. It was less clear to participants in this study whether it might be necessary to transmit those characters back to Facebook in order to make those suggestions.

4.3 Awareness and Privacy Concern

I ran a third Proportional Odds Multinomial Logistic Regression to evaluate the relationship between awareness (perceived likelihood) of automatic data collection and privacy concern. I used *Site* and three of the index variables created from the exploratory factors, described above as regressors. These variables represent participants' perceptions of the likelihood that Google or Facebook can collect First Party Data (*first.party.data*), data from Aggregation Across Sources (*source.aggregation*), or data from Aggregation Across People (*people.aggregation*). The dependent variable was the same privacy concern variable as the previous multinomial regressions: whether participants would be concerned about unwanted access to private information in the hypothetical situation they were exposed to (Yes, Maybe or No). I also included the two continuous controls, Internet Literacy and Privacy Preferences, in the model. The purpose of this analysis was to identify whether a relationship exists between participants' beliefs about how likely it is that their behaviors online are recorded, whether inferences based on that data are possible, and their concern about privacy.

I generated three sets of predicted probabilities from this model to help with interpretation. First, I held the values of all regressors at their means except for *first.party.data*, for which I generated predicted probabilities at 10-point increments between 0 and 100. I did the same for *source.aggregation* and for *people.aggregation*, holding all other regressors at their means. This allows for comparison of the effects of increasing awareness of these three types of information on the predicted probability that a participant would report *Yes*, they would be concerned about unwanted access to private information. Figure 7 depicts these results graphically. Each line in the graph represents one set of predicted probabilities. The predicted probabilities for Facebook and Google are presented separately due to the statistically significant effect of *Site* in this regression. Predicted probabilities of concern are higher for Facebook than for Google.

Figure 7 illustrates that an increase in the perceived likelihood that First Party Data can be collected automatically was associated with a DECREASE in the predicted probability of a participant expressing privacy concern. The more a participant was aware of automatic First Party Data collection, the less concerned he or she was about unwanted access to private information. The open-ended explanations indicated that many participants felt things like what time of day they visit or what links they click on did not need to be kept private. However, as the perceived likelihood of inferences enabled by Source or Person aggregation increase, the predicted probability of of concern about unwanted access to private information also INCREASES. The more a participant believes these inferences are possible, the more likely he or she was to express privacy concern.

	Estimate	*Odds Ratio*	*Std. Error*
Site: Google	-0.498*	0.608	0.197
first.party.data	-0.007	0.993	0.006
source.aggregation	0.011**	1.011	0.004
people.aggregation	0.007*	1.007	0.004
internet.literacy	-0.047	0.955	0.103
privacy.prefs	0.930***	2.535	0.165

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: Coefficients for the Proportional Odds Mulitnomial Logistic Regression. The dependent variable represents participants' level of concern over unwanted access to private information. The Baseline condition is Facebook. AIC is 1364.8; McFadden's Pseudo-R^2 is 0.0471.

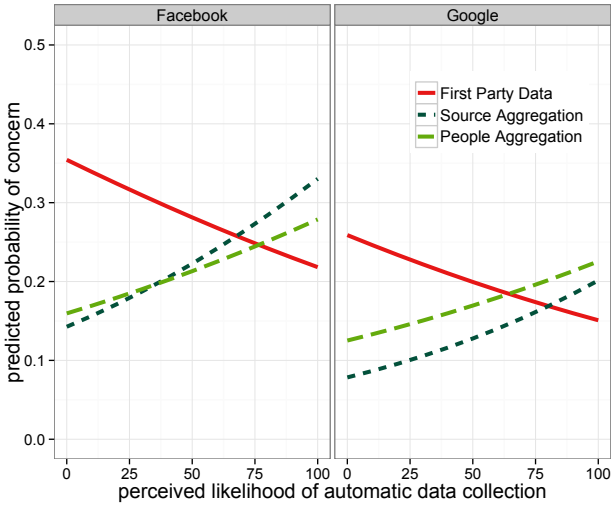


Figure 7: Predicted probabilities from the model in Table 3. The x-axis represents participants' perceived likelihood that Facebook or Google can automatically collect data about them, and the y-axis represents predicted probability of answering *Yes* to the question about privacy concern.

5. DISCUSSION

The data collection technologies and algorithms supporting personalization and behavioral advertising have developed quickly and invisibly, and for web users it is increasingly hard to avoid this "surveillance by algorithm"[2]. Using the web discloses information simply by virtue of interacting with web pages, and then once the information is out of users' control, they have little choice but to trust companies and other people to protect the information the same way they would [22]. Not every user will feel great risk of harm by having their sexual orientation inferred. But, some users might want to keep information like this private, and they presently have no control over it if they want to use the web. They cannot effectively manage that boundary without withdrawing from the Internet altogether. This paper shows that users' perceptions about what unwanted access looks like have very little resemblance to the actual ability of personalization and advertising algorithms to make inferences about them, and this problem will only grow as networked sensors (and the efficiencies and conveniences they provide) become more integrated in our daily activities.

[2] https://www.schneier.com/blog/archives/2014/03/surveillance_by.html

The high-level question that motivated this research project is, when do users currently feel like their actions online are being observed—not necessarily by other people, but recorded by the system—and aggregated to make inferences about them? This is an important question, because if we know more about what situational characteristics are already cause for concern from the user's perspective, we might be able to create systems that are more transparent in the right places about what the system can infer about them.

The results of this study reflect the general trend that participants who were asked about Facebook were more likely to report concern about unwanted access than participants asked about Google. After controlling for participants' level of Internet Literacy and Privacy Preferences, participants were most likely to express concern in the Facebook:Ad conditions, while participants in the Google:Link:Low Sensitivity condition were the least likely group to express concern in the entire study. There is also some evidence in participants' explanations to suggest that they believed clicking a link in Facebook discloses information about them, but that if the same action is part of a Google Search it is not a disclosure. For example, a participant in the Facebook:Link condition wrote, "I hate that facebook knows what im interested in especially when I don't consent it [sic]," indicating that he or she believes Facebook learns about users' interests from what links they click on in the News Feed. In contrast, a participant in the Google:Link condition wrote, "I would not be concerned. I clicked the link and it took me to the place that I wanted" which reflects the perception that links in search results are for navigation only.

Ads in Facebook were more a source of concern for participants than ads in Google, because they perceived that Google ads were associated with search queries (that participants just wouldn't enter if they were sensitive), while Facebook ads were associated with personal characteristics (that participants might not want to reveal). Ads on Facebook contain evidence of aggregation. They're like little windows, not into what the system has collected about users, but into what the system has inferred about them. However, even targeted ads on Google were perceived to only reveal information that the user already gave to Google: the search query. Google may simultaneously provide both a greater feeling of control (over what search terms are entered and what happens when links are clicked), and less feedback that data aggregation is taking place (via the perception that ads are only related to search terms, not profiles).

The main difference between social versus information privacy is the behind-the-scenes aggregation and analysis that is pervasive when interacting with systems, but that does not take place when interacting with other people. The individual bits of information we reveal mean something different, in isolation, than they do as part of a processed aggregate. The invisibility of the infrastructure, from the users' perspective, is both blessing and curse: personalization holds the promise of better usability and access to information, but at the same time the fact that we can't see it makes it harder for us to understand its implications [8].

Most design and policy solutions for privacy issues assume a boundary management model, either by creating mechanisms for specifying what information should be revealed to whom, by providing information about what will be collected and how it will be used and allowing users to opt in or out (notice and choice), or by describing who has rights to ownership and control of data and metadata. The regulatory environment surrounding digital privacy relies on stakeholders to report violations [38], but this is not possible if users cannot tell violations are happening, nor are there laws and mechanisms in place for users to correct mistaken inferences that a system has made about them. Boundary management solutions rely on knowledge and awareness on the part of the user that data is being collected and used.

This study highlights a challenge for privacy research and system design: we must expand our understanding of user perceptions of data aggregation and when feedback about it triggers information privacy concern, so that we might design systems that support better reasoning about when and how systems make inferences that disclose too much. If users are presently unable to connect their behaviors online with the occurrence of unwanted access via inferences made by algorithms, then the current notice and choice practices do not have much chance of working. However, if there are cues in particular situations that users are already picking up on, like ads in Facebook that allow users a glimpse of what the system thinks it knows about them, perhaps the research community can build on these and invent better ways to signal to users what can be inferred rom the data collected about them.

6. ACKNOWLEDGMENTS

Thank you to the BITLab research group at MSU for helpful discussions about this project, and to Paul Rose for assisting with the content analysis. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1217212. The AT&T endowment to the TISM department at MSU also provided support for this project.

7. REFERENCES

[1] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani. Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising. In *SOUPS 2013*, pages 1–16, July 2013.

[2] I. Altman. Privacy: A Conceptual Analysis. *Environment and Behavior*, 8(1):7–29, Mar. 1976.

[3] a. J. Berinsky, G. a. Huber, and G. S. Lenz. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3):351–368, 2012.

[4] J. Bonneau and S. Preibusch. The Privacy Jungle: On the Market for Data Protection in Social Networks. In *Workshop on the Economics of Information Security (WEIS)*, May 2009.

[5] J. T. Child, J. C. Pearson, and S. Petronio. Blogging, Communication, and Privacy Management: Development of the Blogging Privacy Management Measure. *JASIST*, 60(10):217–237, 2009.

[6] S. Das and A. Kramer. Self-Censorship on Facebook. In *ICWSM 2013*, 2013.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting Depression via Social Media. In *ICWSM '13*, July 2013.

[8] R. de Paula, X. Ding, P. Dourish, K. Nies, B. Pillet, D. Redmiles, J. Ren, J. Rode, and R. S. Filho. Two Experiences Designing for Effective Security. In *SOUPS 2005*, pages 25–34, 2005.

[9] D. A. Dillman, J. D. Smyth, and L. M. Christian. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Wiley, Hoboken, NJ, 3 edition, 2009.

[10] C. Duhigg. How Companies Learn Your Secrets. *New York Times*, Feb. 2012.

[11] S. Egelman, A. Oates, and S. Krishnamurthi. Oops, I Did It Again: Mitigating Repeated Access Control Errors on Facebook. In *CHI '11*, pages 2295–2304, 2011.

[12] E. Gilbert. Designing social translucence over social networks. In *CHI '12*, pages 2731–2740, New York, New York, USA, 2012. ACM Press.

[13] M. J. Halter. The stigma of seeking care and depression. *Archives of Psychiatric Nursing*, 18(5):178–184, Oct. 2004.

[14] E. Hargittai and Y. P. Hsieh. Succinct Survey Measures of Web-Use Skills. *Social Science Computer Review*, 30(1):95–107, 2011.

[15] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. *WWW '07*, page 151, 2007.

[16] S. Kairam, M. Brzozowski, D. Huffaker, and E. H. Chi. Talking in Circles: Selective Sharing in Google+. *CHI 2012*, pages 1065–1074, 2012.

[17] A. Korolova. Privacy Violations Using Microtargeted Ads: A Case Study. *Journal of Privacy and Confidentiality*, pages 27–49, 2011.

[18] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805, 2013.

[19] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, Mar. 1977.

[20] E. Litt. Understanding social network site users' privacy tool use. *Computers in Human Behavior*, 29(4):1649–1656, 2013.

[21] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook Privacy Settings: User Expectations vs. Reality. In *IMC 2011*, pages 1–7, 2011.

[22] S. T. Margulis. Three theories of privacy: An overview. In *Privacy Online: Perspectives on Privacy and Self-Disclosure in the Social Web*, pages 9–18. Springer Verlag, 2011.

[23] L. A. Martin, H. W. Neighbors, and D. M. Griffith. The Experience of Symptoms of Depression in Men vs Women: Analysis of the National Comorbidity Survey Replication. *JAMA Psychiatry*, Aug. 2013.

[24] M. a. Moreno, L. a. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, and T. Becker. Feeling bad on Facebook: depression disclosures by college students on a social networking site. *Depression and Anxiety*, 28(6):447–455, 2011.

[25] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless Monster: Exploring the Ecosystem of Web-based Device Fingerprinting. In *IEEE Symposium on Security and Privacy*, pages 1–15, 2013.

[26] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books. Stanford Law Books, 2009.

[27] S. Panjwani and N. Shrivastava. Understanding the Privacy-Personalization Dilemma for Web Search: A User Perspective. In *CHI 2013*, pages 3427–3430, 2013.

[28] K. Purcell, J. Brenner, and L. Rainie. *Search Engine Use 2012*. Pew Research Center's Internet & American Life Project, Washington, D.C., Mar. 2012.

[29] E. Rader, A. Velasquez, K. D. Hales, and H. Kwok. The gap between producer intentions and consumer behavior in social media. In *GROUP '12*. ACM Request Permissions, Oct. 2012.

[30] L. Rainie, S. Kiesler, R. Kang, and M. Madden. *Anonymity, Privacy, and Security Online*. Pew Research Center's Internet & American Life Project, Washington, D.C., Sept. 2013.

[31] S. Sengupta. On Facebook, 'Likes' Become Ads. *New York Times*, May 2012.

[32] A. Sharma and D. Cosley. Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems. In *WWW '13*, pages 1133–1143, 2013.

[33] S. Silfverberg, L. A. Liikkanen, and A. Lampinen. "I'll press Play, but I won't listen": Profile Work in a Music-focused Social Network Service. In *CSCW 2011*, pages 207–216, 2011.

[34] N. Singer. You for Sale: Mapping, and Sharing, the Consumer Genome. *New York Times*, June 2012.

[35] M. Sleeper, R. Balebako, and S. Das. The Post that Wasn't: Exploring Self-Censorship on Facebook. In *CSCW '10*, pages 793–802, 2013.

[36] H. J. Smith, T. Dinev, and H. Xu. Information Privacy Research: An Interdisciplinary Review. *MISQ*, 35(4):989–1016, Nov. 2011.

[37] H. J. Smith, S. J. Milberg, and S. J. Burke. Information Privacy: Measuring Individuals' Concerns about Organizational Practices. *MISQ*, 20(2):167–196, 1996.

[38] D. J. Solove. Introduction: Privacy self-management and the consent dilemma. *126 Harvard Law Review*, pages 1880–1903, 2013.

[39] F. Stutzman and W. Hartzog. Boundary Regulation in Social Media. In *CSCW 2012*, pages 769–778, 2012.

[40] E. Toch, Y. Wang, and L. F. Cranor. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2):203–220, 2012.

[41] B. Ur, P. L. Leon, L. F. Cranor, R. Shay, and Y. Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *SOUPS '12*, 2012.

[42] A. F. Westin. Social and Political Dimensions of Privacy. *Journal of Social Issues*, 59(2):431–453, Apr. 2003.

[43] C. E. Wills and C. Tatar. Understanding What They Do with What They Know. In *WPES 2012*, pages 13–18, 2012.

APPENDIX

A. SURVEY QUESTIONS

Data collected: May 10 – 16, 2013

Sample: 701 Amazon Mechanical Turk workers who were 18 or older, had a 95% or higher approval rating after completing at least 500 tasks, and reported in the screening questionnaire that they visited both Facebook and Google Search at least weekly.

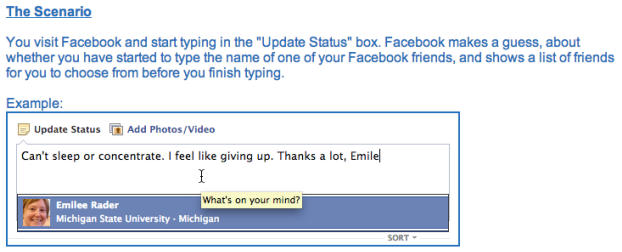
A.1 The Scenarios

In this section of the survey, you will be shown an example of a scenario people often encounter when using Facebook or Google Search.

As you read the scenario, please think about what it would be like for you to experience something like it.

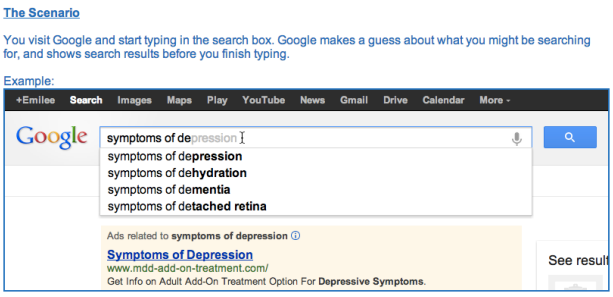
Autocomplete, Facebook, Non-Sensitive.

Autocomplete, Facebook, Sensitive.



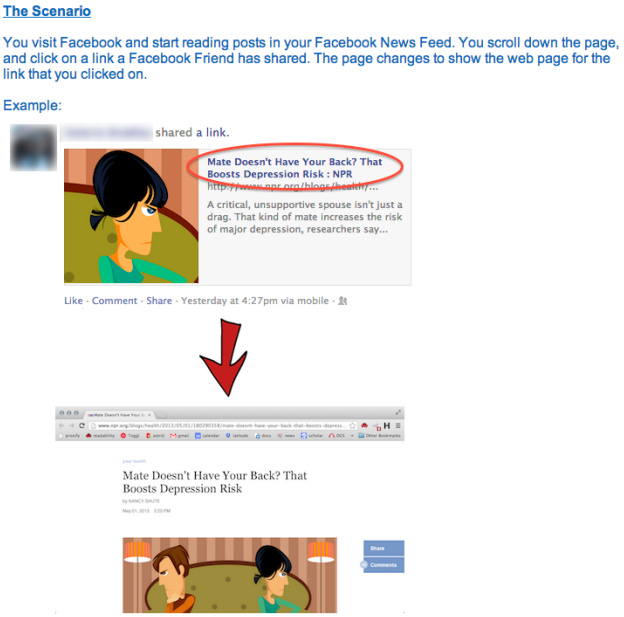
Autocomplete, Google, Non-Sensitive.

Autocomplete, Google, Sensitive.



Link, Facebook, Non-Sensitive.

Link, Facebook, Sensitive.

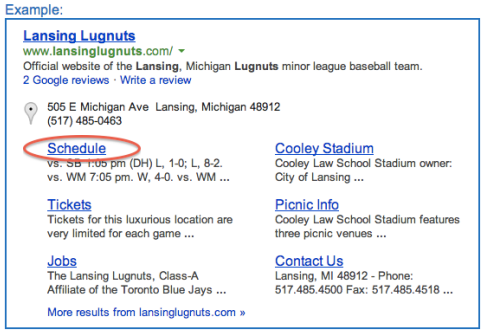


Link, Google, Non-Sensitive.

The Scenario

You are viewing the results of a Google search. You scroll down the page to find the information you are looking for, and then click on the link. The page changes to display the web page for the link that you clicked on.

Example:

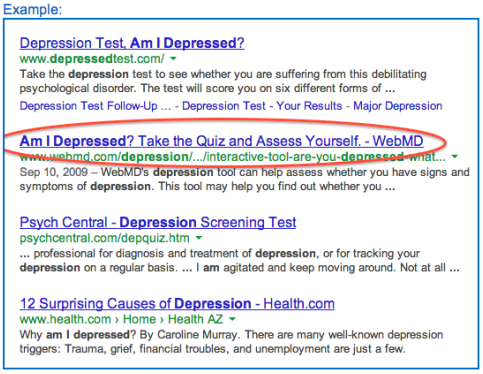


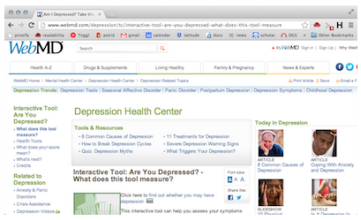
Link, Google, Sensitive.

The Scenario

You are viewing the results of a Google search. You scroll down the page to find the information you are looking for, and then click on the link. The page changes to show the web page for the link that you clicked on.

Example:



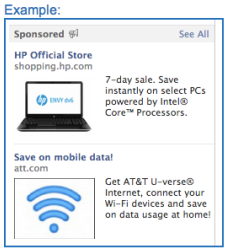


Ad, Facebook, Non-Sensitive.

The Scenario

You are viewing posts in your Facebook News Feed. As scroll down the page, reading posts made by Facebook friends, you notice ads displayed on the right side of the screen.

Example:

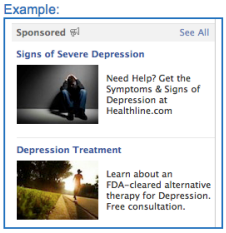


Ad, Facebook, Sensitive.

The Scenario

You are viewing posts in your Facebook News Feed. As scroll down the page, reading posts made by Facebook friends, you notice ads displayed on the right side of the screen.

Example:

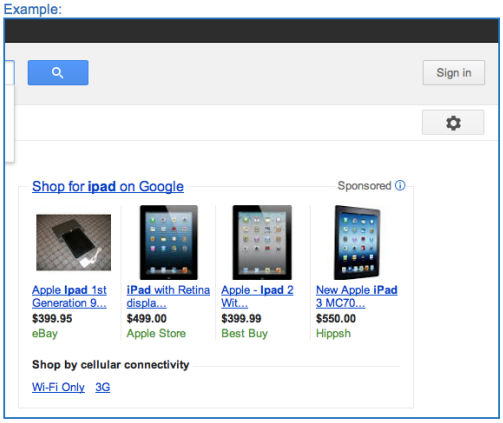


Ad, Google, Non-Sensitive.

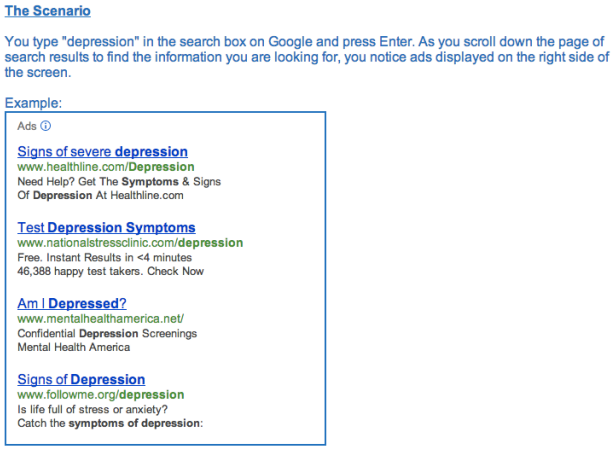
The Scenario

You type "ipad" in the search box on Google and press Enter. As you scroll down the page of search results to find the information you are looking for, you notice ads displayed on the right side of the screen.

Example:



Ad, Google, Sensitive.



A.2 Concern

Q1 Would you be concerned about unwanted access to private information about you in this scenario? (Yes=151, Maybe=173, No=377)

Q2 Please explain your answer to the previous question. (open-ended)

Q3 What would you tell someone else about how to control private information in the above scenario? Please describe what you would say, below. (open-ended)

A.3 Information Types

AWARENESS How likely do you think it is that [Google | Facebook] can AUTOMATICALLY record each of the following types of information about you? Please indicate below how likely you believe each example is on a scale from 0-100, where 0 means Unlikely, and 100 means Likely.

M	*SD*	
92.0	15.6	what time of day you visit [Google | Facebook]
84.9	19.9	your physical location when using [Google | Facebook] on a mobile device
65.0	32.9	what you are typing in the [search | Post or Comment] box before you submit the [search terms | post]
80.0	25.5	how much time you spend reading [Google | Facebook] status updates
71.8	30.6	what kind of computer you are using when you visit [Google | Facebook]
81.2	23.9	your physical location when using [Google | Facebook] on a computer
19.7	25.1	what you purchase from a brick-and-mortar store
60.1	31.2	what you purchase from online shopping websites
69.6	29.8	what websites you visit most often
69.5	30.5	which people you communicate with online most often
50.8	32.7	your political party affiliation
93.2	13.9	how often you visit [Google | Facebook]
50.6	34.7	your sexual orientation
19.1	24.0	what the desktop image on your computer looks like
71.1	29.0	which online retailers (e.g. Amazon.com) you visit most often
91.0	16.2	what links you click on in your [Google search results pages | Facebook news feed]

A.4 Privacy Preferences

PRIVACY PREFS Here are some statements about personal information. From the standpoint of personal privacy, please indicate how much you agree or disagree with each statement below. [Strongly Disagree (1) Disagree (2) Neutral (3) Agree (4) Strongly Agree (5)]

M	*SD*	
4.36	0.82	If I think that information I posted to Facebook really looks too private, I might delete it.
4.08	4.27	I don't post to Facebook about certain topics because I worry who has access.
2.93	1.20	I use shorthand (e.g., pseudonyms or limited details) when discussing sensitive information on Facebook so others have limited access to know my personal information.
4.03	0.90	I like my Facebook status updates to be long and detailed. REVERSE CODE
4.17	0.95	I like to discuss work concerns on Facebook. REVERSE CODE
4.36	0.81	I have limited the personal information that I post to Facebook.
3.81	1.05	When I face challenges in my life, I feel comfortable talking about them on Facebook. REVERSE CODE
3.71	1.05	When I see intimate details about someone else on Facebook, I feel like I should keep their information private.
4.33	0.88	When people give personal information to a company for some reason, the company should never use the information for any other reason.
3.99	0.96	It usually bothers me when companies ask me for personal information.
4.42	0.90	Companies should never sell the personal information in their computer databases to other companies.
3.83	1.01	I'm concerned that companies are collecting too much personal information about me.

A.5 Scenario Realism

AUTOCOMPLETE only Search engines and social media websites can make a guess about what you are about to type, while you are typing, and provide you a list of suggestions – like in the scenario displayed at the beginning of this survey. Have you ever used a website that has this "autocomplete" functionality? [Yes=227, No=6]

LINK only Search engines and social media websites provide links (URLs) to content on other websites containing information that is interesting, entertaining, etc. – like in the scenario displayed at the beginning of this survey. Have you ever clicked on a link in a search engine or social media website that took you to content on some other website? [Yes=224, No=10]

AD only Search engines and social media websites can display personalized or "targeted" advertising – like in the scenario displayed at the beginning of this survey. Have you ever noticed "targeted" advertising when surfing the web? [Yes=228, No=6]

A.6 Internet Literacy and Experience

INTERNET LITERACY How familiar are you with the following Internet-related terms? Please rate your familiarity with each term below from None (no understanding) to Full (full understanding): [None (1) Little (2) Some (3) Good (4) Full (5)]

	None	Little	Some	Good	Full
Wiki	1	23	52	187	438
Netiquette	129	61	121	175	215
Phishing	18	48	92	225	318
Bookmark	4	7	22	146	522
Cache	11	44	137	236	273
SSL	171	159	136	113	122
AJAX	409	131	83	37	41
Filtibly (FAKE WORD)	587	85	29	0	0

E1 Have you ever worked in a Òhigh techÓ job such as computer programming, IT, or computer networking? [Yes=115, No=586]

E2 How often do you visit Facebook?

Once a Week or less	6
2-3 Times a Week	88
Daily	246
Many times per day	361

E3 How often do you search the web using Google? [Once a Week or less, 2-3 Times a Week, Daily, Many times per day]

Once a Week or less	1
2-3 Times a Week	15
Daily	137
Many times per day	548

E4 Do you use ad blocking software when you browse the web? [Yes=536, No=144, Don't Know=21]

E5 Have you ever had one of the following experiences? Please check all that apply:

No	Yes	
89	612	Received a phishing message or other scam email
34	667	Warning in a web browser that says "This site may harm your computer"
57	644	Unwanted popup windows
154	547	Computer had a virus
646	55	Someone broke in or "hacked" the computer
503	198	Stranger used your credit card number without your knowledge or permission
687	14	Identity theft more serious than use of your credit card number without permission
691	10	None of the above

A.7 Demographics

D1 How old are you? Please write your answer here: [*M*=30.2, *SD*=9.22]

D2 What is the last grade or class you completed in school?

0	None, or grades 1-8
2	High school incomplete (grades 9-11)
71	High school graduate (grade 12, GED certificate)
20	Technical, vocational school AFTER high school
285	Some college, no 4-year degree
241	College graduate (B.S., B.A., 4-year degree)
27	Post-graduate
3	Other
0	I Don't Know

D3 What is your gender? [Man=398, Woman=297, Prefer not to answer=6]

D4 What is your race?

American Indian or Alaska Native	4
Asian or Pacific Islander	63
Black or African-American	41
Hispanic or Latino	26
White	560
Other	7

D5 Which of the following BEST describes the place where you now live?

A large city	155
A suburb near a large city	256
A small city or town	211
A rural area	78
Other	0
Don't know	1

D6 Most people see themselves as belonging to a particular class. Please indicate below which social class you would say you belong to:

Lower class	41
Working class	173
Lower middle class	141
Middle class	276
Upper middle class	69
Upper class	1
Other	0

D7 Are you now employed full-time, part-time, retired, or are you not employed for pay?

Employed full-time	310
Employed part-time	94
Retired	6
Not employed for pay	77
Self-employed	85
Disabled	11
Student	104
Other	14

B. CONTENT ANALYSIS

Respondents were asked to explain why they answered (Yes, Maybe, or No) to a question that asked, "Would you be concerned about unwanted access to private information about you in this scenario?"

The purpose of this coding scheme is to differentiate between two potential themes that appeared in many respondents answers. These themes are informed by the distinction in the literature between "social" privacy – or control over information in relation to other people, and "informational" privacy, or control over information in relation to technologies, organizations or the government.

Each answer should be coded "INFO", "SOCIAL" or "NEITHER".

Step 1. Determine whether the response contains an explicit reference to a potential third party accessing/obtaining information related to the respondent.

If the answer contains no clear reference to a third party, or does not implicate accessing/obtaining respondent info, or does not provide evidence that the coder can use to tell whether the third party access is "social" or "informational", code as NEITHER. Otherwise, proceed to Step 2

In general, responses with ambiguous pronouns without an explicit referent (e.g. "they", "them", "it") should be coded as NEITHER, because without more information from the respondent, it is impossible to tell whether the referent is a person, organization, government, or website. For example, "Really depends on exactly what kind of information they gathered. I am OK with just basic information".

Likewise, the presence of passive voice (e.g. "Private information is being read from my posts"), should be coded as NEITHER, because these responses typically do NOT constitute an explicit reference that allows the coder to differentiate who or what the third party is.

However, there are exceptions to the above. To proceed to Step 2 with a response that contains ambiguous pronouns or passive voice, the response must contain some other evidence that allows the coder to determine whether the potential for unwanted access is SOCIAL- or INFO-related.

This evidence often comes in the form of mentioning ads, IP addresses, databases, or some other technology or feature as if it is involved in information collection, access, or processing. For example, "It would really depend on what kind of information. Not much I can do about them using my IP address to localize the type of ad"; or, "I'm aware that certain things about me are known and will be used to select ads, and I don't mind that".

Step 2. Determine whether the explicit reference to third party access in the response includes evidence that the third party is a human being, or a group of people.

This could include language like "other people", "employers", "friends", "others", "anyone", etc. Pronouns such as "it" and "they" should NOT be treated as SOCIAL, unless the referent is present in the response. If the answer contains evidence that the third party is clearly a person or group of people, code as SOCIAL. If not, code as INFO.

Some answers might legitimately contain references to both people and organizations, governments, or websites. In these cases, try to determine from the response which aspect, SOCIAL or INFO, is causing more concern for the respondent. If it is not clear, code as SOCIAL. Example: "I wouldn't be concerned because even if google is keeping track of what all of their subscribers are looking up, there are so many people in the world that the chances of anyone looking at my individually are slim to none."

B.1 Examples, Site:Code:Concern

Facebook:INFO:Yes.

- I do not feel that ANYTHING that I say on my facebook account is private. It makes me feel strange when a computer is second guesing me before I finish typing.
- It's never comfortable for ad companies to have private information about me.

Facebook:INFO:No.

- I am posting a facebook status on facebook. I don't mind that facebook is guessing who I might be tagging in my facebook status post. All that information can be found on facebook.
- The ads seem random to me and doesn't have anything to do with me.

Google:INFO:Yes.

- I don't believe search information should be logged and associated to persons.
- Most people know that search engines, ESPECIALLY Google, collect all sorts of information about people and then pass it on to the government.

Google:INFO:No.

- I don't care if google knows what I search. I have no secrets.
- The ads are only coming up based on my search. The ads could be helpful.

Facebook:SOCIAL:Yes.

- I'm not sure I want people to know what website's I have been to.
- I am very concerned about my privacy anyway, especially when it comes to things shared on Facebook and other social networks.

Facebook:SOCIAL:No.

- Because, I just clicked on the link. I only would be concern if facebook announced on the news feed that I read the article.
- Because no one else sees me typing in the box and I already know who my friends are, can see my friends list, etc.

Google:SOCIAL:Yes.

- I would be concerned that someone could find out my search for depression by checking my Google search history, and that they keep a record of that when they display ads to me.
- I hate when previous searches pop up while someone is browsing my computer.

Google:SOCIAL:No.

- I am not sure how my provacy would be jeopardized in this scenario. Even if it were, I don't think I'd be concerned if someone were to find out I was searching for help with depression.
- Those ads are automatically displayed to anyone who enters in a particular search term. They don't have anything to do with me individually. I don't see any indications that any information was revealed to the people who placed the ads.

Too Much Choice: End-User Privacy Decisions in the Context of Choice Proliferation

Stefan Korff
Westfälische Wilhelms-Universität Münster
Department of Information Systems
Leonardo-Campus 3, 48149 Münster, Germany
stefan.korff@uni-muenster.de

Rainer Böhme
Westfälische Wilhelms-Universität Münster
Department of Information Systems
Leonardo-Campus 3, 48149 Münster, Germany
rainer.boehme@uni-muenster.de

ABSTRACT

Choice proliferation, a research stream in psychology, studies adverse effects of human decision-making as the number of options to choose from increases. We test if these effects can be elicited in a privacy context. Decision field theory suggests two factors that potentially affect end-users' reflection of disclosure decisions: (1) choice amount, which we test by changing the number of checkboxes in a privacy settings dialog; and (2) choice structure, tested by varying the sensitivity of personal data items which are jointly controlled by each checkbox. We test both factors in a quantitative 2×2 between-subject experiment with stimuli calibrated in a pre-study with 60 respondents. In the main experiment, 112 German-speaking university students were asked to enter personal data into an ostensible business networking website and decide if and with whom it should be shared. Using an established item battery, we find that participants who are confronted with a larger amount of privacy options subsequently report more negative feelings, experience more regret, and are less satisfied with the choices made. We observe a similar tendency, albeit weaker and statistically insignificant in our small sample, for the complexity of the choice structure if the number of options remains constant.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues, Privacy

General Terms

Human Factors

Keywords

Privacy, Choice Proliferation, Experiment

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

1. INTRODUCTION

Proliferation of choice is characteristic for post-industrial societies. It can refer to the number of decisions consumers are asked to make everyday and the number of alternatives to choose from for each decision. Choice proliferation is arguably driven by competition, product and service differentiation, technology-enabled mass customization, and the positive psychological effects inherent to choice [47].

Robust empirical evidence suggests that the provision of choice increases intrinsic motivation, perceived control, and life satisfaction [6]. Past decisions are also reflected on an psychological level where they may cause positive emotional states like satisfaction, happiness; but also negative states like regret, dissatisfaction, and indisposition [36]. A research stream in psychology believes that positive and negative emotional states in a decision-making process are determined by the amount of available options [51, 29]. Moreover, researchers in this field suspect that the accumulation of decision-making tasks is a reason for various negative psychological long-term effects including serious mental diseases, like clinical depression [53].

Choice also plays a key role in the domains of privacy and human-computer interaction. The positive notion of privacy as control over the collection and use of personal data [63, for example] suggests that more choice on information disclosure and sharing decisions is always better. This and the temptation to shift liability, encourages service providers to design more and more granular panels for privacy settings, thereby delegating more privacy decision to the end-user. For instance, Facebook has recently softened the default privacy settings for teenagers by adding more options to put the "decision to share in teens' hands" [33]. The number of end-user decisions in the privacy space is further inflated by legal obligations to inform consumers about the purpose of personal data collection and to request explicit consent [14].

Behavioral aspects of end-user privacy decisions are increasingly being studied. However, psychological side-effects of data disclosure and sharing decisions on individuals have rarely been addressed. This work tries to close this gap. It draws on the seminal choice proliferation literature and connects it with privacy research to explain why and how the amount of choice impacts end-users' privacy decisions and causes inherent psychological side-effects. We propose a decision-making model, based on elements of decision field theory, to derive testable choice scenarios (conditions) that are hypothesized to effect the users' decision-making process and its emotional reflection. In the main study (Study 2),

we asked 112 participants to create a CV-like user profile in a guided process and let them subsequently decide which members of an ostensible business networking website they permit to view (parts of) their profile. The conditions modify the amount and structure of the information sharing decision. The key stimuli of the four conditions in our between-subject experimental design have been validated in a preceding quantitative study (Study 1).

By using scales established in consumer psychology and marketing research, we measure the participants' emotional reflections of the choice process. Our results are twofold: first, we show that an increase in choice amount correlates with more negative feelings towards past decisions; second, we identify behavioral attributes that may explain individual differences in the effect size.

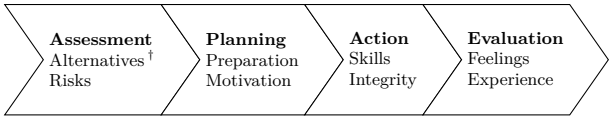
The remainder of this paper is organized as follows. Section 2 develops a theoretical framework. Sections 3 presents research questions and the empirical approach. The pre-study and the main experiment are reported in Sections 4 and 5, respectively. Section 6 discusses limitations and the final Section 7 summarizes and concludes with an outlook.

2. THEORETICAL BACKGROUND

Developing robust hypotheses on the relation between the amount of personal data sharing options and emotional reflections of the decision requires some solid theory. In this section, we adapt a general model of human decision-making from behavioral sciences to the specific domain of information privacy and, where appropriate, establish relations to prior experimental research of end-user privacy behavior.

2.1 Models of Human Decision-Making

There exist multiple psychological and cognitive frameworks that model conscious, rational or partly rational decision-making processes. The *Rubicon model*[1], for instance, describes a decision as a four-step process [28]. Figure 1 shows an adapted version. According to the model, a decision process starts with the assessment phase. The individual deliberates between possible alternatives by considering past experiences, knowledge, perceived risks, and valences. In the subsequent planning phase, necessary preconditions for the required actions are collected and elaborated. If the preferred alternative is selected and the necessary preparation is completed, the decision is translated into a sequence of actions. A decision-making process closes with an evaluation. The experience made and the fulfillment of intentions are reflected. The evaluation results are memorized as feelings (regret, satisfaction, etc.) and knowledge. They can be recalled for subsequent decisions.



† Structure and amount (our experimental factors)

Figure 1: Adapted version of the "Rubicon model" of action phases, cf. [28].

[1]There are various other famous choice models, like the preference trees [61] or elimination by aspects [60].

The assessment phase is an essential step in the decision-making process and there are various models that describe this phase in more detail. One of them is the *decision field theory* (DFT) [13]. DFT is a cognitive stochastic decision-making model which describes the process of deliberating between choice alternatives over time. The model assigns a payoff function to each alternative and explicitly assumes that humans accumulate the valence of each alternative over time. The preferred option can change various times during the accumulation process. This is captured in the model by letting payoffs fluctuate along stochastic processes. At any given point in time, the preferred alternative is the one with the highest accumulated payoff. The theory defines three thresholds, any of which, if exceeded, causes a termination of the decision process.

(1) The *decision boundary* defines the minimum amount of accumulated payoff required by an alternative to be considered as the final, most preferable outcome.

(2) The *time threshold* sets a maximum time for the decision process.

(3) The *preference change threshold* defines the maximum number of preference changes.

Upon termination, the decision maker selects the then preferred option or defers the decision. Decisions not terminated by the decision boundary amplify possible negative reflections of the decision process [30].

The frequency of preference changes and the time needed to make a decision are determined by the amount and distribution of the available alternatives. Concepts useful to model the size and the structure of choice alternatives are *density* and *entropy* [30, 22]. The terms originate from mathematics and information theory and were adopted in consumer psychology and marketing research to describe product assortments as inputs to a decision process. ("Options" and "products" are synonyms for the original term "alternatives" in the Rubicon model.) According to the density model, products are points in a high-dimensional attribute space. In a dense structure, the attribute values of each option lie closely together. A scattered structure is characterized by an increased number of extreme values and a larger distance between the attribute values of each option. While density requires numeric attribute values, entropy is applicable to numerical and categorial attribute values alike. It is calculated per attribute using Shannon's theory [54]. For a fixed number of options, an increase in the number of attribute values leads to a higher entropy.

The effectiveness of decision-making and in particular the assessment phase is determined by the decision strategy pursued. The set of strategies individuals apply to solve the same decision problem can be completely heterogeneous in many aspects [43, 24]. Each strategy implies a different configuration of the decision boundaries [17, 30]. A person who tries to optimize the benefit of the decision outcome, from now on called *optimizer*, sets a higher decision boundary and accepts more changes in preferences and more time to reach this boundary than a person who is not very interested in the decision quality, in the following called *satisficer* [55]. Assuming the same cognitive capabilities for both types of decision makers, an optimizer needs more resources than a satisficer. With limited resources it becomes harder to reach

the decision boundary if the frequency of preference changes is high or time is scarce.

We conclude from this model: first, more options require a decision maker to assess more information, which requires more time and resources; second, if the option structure is dense or has high entropy, the number of preference changes is higher than in the case of a scattered option structure. That's because extreme non-fitting options can be easily sorted out at the beginning of the assessment phase. Both features affect optimizers more than satisficers.

2.2 Mechanisms of Choice Proliferation

Choice proliferation subsumes the increasing number of decisions and the growing number of available alternatives in the assessment phase of a decision process. Choice proliferation is studied by researchers originating from psychology, marketing, consumer research [50], and occasionally computer science [41]. As a result, there is no unified terminology and many terms exist to emphasize the negative consequences of choice proliferation, such as "tyranny of too much choice", "choice overload", or the "too much choice" (TMC) effect. We use the latter acronym to refer to the phenomenon.

On the upside, one should not forget that more choice primarily goes along with more freedom and autonomy [23]. It allows people to fulfill their individual needs and express their preferred way of living [48]. Choice further enables the exercise of control over the environment and prevents people from feeling helpless [15, 35, 46]. The benefits of having choice are therefore essential for human well-being.

On the downside, as indicated by the decision-making models, an increase in choice amount requires a decision maker to process more information and make more trade-offs. It increases the frequency of preference changes and time needed. As a consequence, individuals might fail to reach a decision boundary. Furthermore, more options imply more parameters to handle in order to maximize the decision output. This raises the decision boundary while the resources remain fixed. Both effects may trigger negative reflections of the decision-making process [51]. Researchers have two explanations for this link: first, the amount of options might exceed the cognitive capabilities of maintaining control, which provokes helplessness [53]. Second, more choice fuels the expectations to find the perfect satisfaction of needs, which, if not met, leads to the experience of regret, dissatisfaction, and disenchantment [8].

A general assumption of TMC studies is that effects triggered in the assessment phase materialize in the evaluation phase of a decision process (cf. Fig. 1). For example, in the domain of marketing, the amount of options, as part of the assessment phase, has a strong impact on the subsequent planing and action phase, i. e., consumers' purchase behavior [56]. Iyengar and Lepper's "jam study" [29] is generally considered as seminal for this field. They placed a tasting booth for different jams in a supermarket. All jams were of the same brand and could be purchased in the store. Customers in the control group of the between-subject experimental design were invited to sample no more than two jams from an array of six different flavors. Customers in the experimental group were allowed to sample two jams among 24 different flavors. The tasting booth which offered a larger amount of options attracted more people than the smaller one. But, customers in the control group were considerably more likely to purchase the product than customers of the experimental group. The researchers relate the unwillingness to purchase to choice deferral, which is believed to be an indicator for the TMC effect.

More than 40 follow-up studies provide empirical evidence on the TMC effect (see [50] for a survey). The dominant approach to simulate choice and its proliferation is to ask human subjects to choose one object out of a set of comparable options. The size of the set is varied between experimental conditions. Studies that successfully reveal the choice overload effect find a correlation between negative psychological effects (regret, dissatisfaction), measured with standardized instruments, and the amount of options. More recent research includes additional factors as control variables, such as time or the option structure. Apparently, not only the amount of options, but the overall complexity of the decision problem including option structure [22], time [27], etc. causes the TMC effect.

2.3 Choice and Privacy

Privacy is a multi-faceted concept and there is no universal consensus on its dimensions [57]. However, most scholars agree that privacy never implies absolute protection and emphasize the freedom of the individual to give up some privacy, for example by overriding safe defaults. Exercising this option implies choice, which raises the question on how this choice should be presented to end-users.

The term "choice architecture" has been coined by Thaler and Sunstein [59], who illustrate that the way how choice is presented can profoundly impact decision outcomes. In the context of privacy research, the term is used to describe the visual and logical presentation and composition of options that allow individuals to manage privacy-related tasks [19]. For example, the distinction between opt-in and opt-out policies [11] or the framing of consent dialogs [10] can be interpreted as instances of choice architecture. While privacy activists voice concerns about the inherent possibility of manipulation towards laxer sharing of personal data, empirical privacy research suggests that choice architecture can also be leveraged to nudge (i. e., manipulate) consumer behavior in the opposite direction, or at least to reach more conscious and thus fairer privacy decisions [2, 7].

There is a small but growing body of literature discussing the preservation of privacy under psychological constraints, comparable to the proliferation of choice. For instance, Böhme and Grossklags [9] comment on the trend to delegate all kinds of security decisions (security warnings, end-user license agreements, privacy notices and consent forms) from vendors to end-users. Since human decision capacity is scarce, only the most important decisions should be handled by the end-users. But individual vendors have little incentive to unilaterally suppress less important decision requests, such as take-it-or-leave-it decisions at install time, leading to habituated responses (clicking dialogs away) and a discrepancy between ostensible and actual control. Brandimarte et al. [12] analyze what they call *control paradox of personal information* empirically: If users have more control over the flow of personal data, they tend to disclose more sensitive information. Apparently, the perceived increase in control outweighs concerns regarding the subsequent access and usage of personal information. Other researchers have experimented with different granularity of privacy control options in a mobile location sharing scenario [32, 58]. All

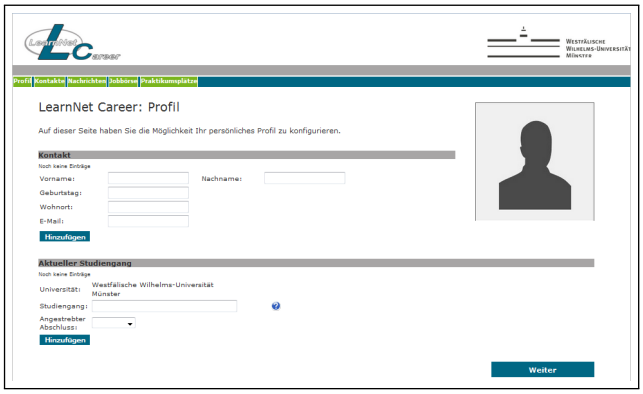


Figure 2: Profile page of the business networking website. (Original screenshot from Study 2)

studies demonstrate in a data disclosure context that the outcome of a decision process depends on the structure of the available options. Our research differs in that the dependent variable is not the decision outcome (disclosure), but, in line with the TMC tradition, psychological reflections in the evaluation phase.

What remains is to reason about the experimental factor (choice amount) in a privacy context. Practical online privacy management builds on a number of mechanisms that allow individuals to actively manage data disclosure and access permissions: consent dialogs [10], privacy settings [26], or the data entry fields where the actual disclosure happens [44]. A very general way of looking at data usage permissions is the access control matrix (ACM) [34]. An ACM consists of a set of objects O (originally: system resources, files, processes), protected by access rights, and a set of subjects S (users or processes operating on their behalf), who can be granted those rights. In a privacy context, the resource correspond to personal data items and the subjects can be recipients or purposes. The amount of options is determined by the number of O (objects) and S (subjects). An $O \times S$ matrix requires the user to decide $O \times S$ times whether a given piece of information should be shared with the given subject or not.

3. EMPIRICAL APPROACH

Our general goal is to empirically investigate possible TMC effects for end-user privacy decisions. To this end, we derive specific research questions from the presented theory (Sect. 3.1), develop a plausible scenario where TMC effects may appear (Sect. 3.2) and can be measured with specifically tailored stimuli (Sect. 3.3) in a between-subject experimental design. We run a pre-study to calibrate the stimuli (Study 1, Sect. 4) and collect empirical evidence to answer the research questions in the main study (Study 2, Sect. 5).

3.1 Research Questions

The designated approach is to design an ACM for personal data sharing permissions, as introduced in Section 2.3, to simulate different disclosure decision scenarios. An ACM can be manipulated pretty flexibly to adjust the amount of choice ($O \times S$). We formulate our first research question (RQ) accordingly:

Research Question 1. *How does the number of options presented in an access control matrix for personal data sharing permissions affect the attitudes towards the decisions in the reflection phase?*

As outlined in Section 2, individuals reflect a decision more negatively not only for the increased choice amount, but also for the inherent complexity of the decision. We understand the complexity as a latent factor moderated by the structure of the presented options. In the context of privacy, a suitable actuator for the decision complexity—independent of the number of options—is the perceived sensitivity of data items bundled together as objects in the ACM. More specifically, we assume persons who must decide if a set of data items with similar sensitivity should be disclosed or not face a less complex decision than those who are presented with more heterogeneous sets. Therefore:

Research Question 2. *How does the complexity of a data sharing decision, represented by the grouping of items of more or less similar sensitivity, affect the attitudes towards the decisions in the reflection phase?*

The attitude towards the decision in RQ 1 and RQ 2 is operationalized by two measurement scales. The TMC scale combines established items which measure emotional effects (satisfaction, regret, feeling overwhelmed) as an indicator for having too much choice. Second, the combined items of our perceived comfort, risk and trustworthiness (PCRT) scale are designed to capture how the participants feel while interacting with a website, specifically. Clearly, the PCRT scale is more exploratory. It has not been used in TMC studies before.

Also on the exploratory side and as a follow-up question, we are interested to which extent character traits may cause people to be more or less prone to the TMC effect.

Research Question 3. *Can we find individual differences moderating the TMC effect in privacy decisions?*

In particular, we are interested in how privacy concerns (PC scale) and the participants' generally pursued decision strategy (MAX scale) affect ratings on the TMC and PCRT scales. The PC scale combines different established items used in privacy research. With the help of the MAX scale, originally developed by Schwartz et al. in [52], we can classify participants into satisficers or optimizers. (All scales are further described in Sect. 5.1.)

3.2 Scenario

The main difficulty of adopting TMC studies from the domain of consumer and marketing research to privacy is that attributes of data sharing options are much more abstract than properties of tangible goods. For classical goods, consumers have formed expectations, often based on experience, and they have a clear and largely homogenous idea of the value. By contrast, the costs and benefits of privacy options are rarely monetary and therefore less salient and hard to assess and compare. In general, privacy preferences, attitudes and behavior alike, differ substantially between individuals [3]. Simulating an information disclosure situation which is perceived as an actual decision process more or less uniformly by all participants of an experimental study, while maintaining external validity and satisfying practical and ethical constraints, turned out to be quite challenging.

We follow [39] and choose a job market scenario for our TMC experiment. We created a business networking website

inspired by existing services like LinkedIn or Xing[2]. Unlike popular open networks, our service was described as exclusive offer to students and graduates of one large German university who can use the platform to get in touch with potential employers. As a special feature, the service might authentically signal grades and recommendations from the university to the job market. To support this cover story, we called the service "Learnnet Career", alluding to the name of the Moodle-based e-learning and course management platform of the university. Study participants were invited to serve as beta-testers of a prototype of the new platform. They were asked to log in with their campus account and complete a CV-like profile (see Fig. 2). Tooltip examples next to the entry dialogs as well as the assurance that temporary information can be corrected and completed with details later (e. g., exact dates), helped to reduce the barriers of entering valid information. Data available in the campus directory was offered for direct import into the profile.

3.3 Experimental Conditions

After completion of the profile, participants were asked to configure data sharing permissions with an ACM. All defaults were set to not sharing and the participants had to opt in by checking the corresponding boxes. We varied the size and the structure of the ACM between four experimental conditions to elicit the TMC effect.

Privacy Settings			
Profile Information (Objects)	Members (Subjects)		
	Fellow Students	All Employers	All network members
Name Surname Age	☐	☐	☐
Job Experience Education	☐	☐	☐
Relationship status Political Interests	☐	☐	☐

(a) Homogeneous object structure.

Privacy Settings			
Profile Information (Objects)	Members (Subjects)		
	Fellow Students	All Employers	All network members
Name Surname Relationship status	☐	☐	☐
Job Experience Age	☐	☐	☐
Education Political Interests	☐	☐	☐

(b) Heterogeneous object structure.

Figure 3: Modifying the object structure in an ACM of a business networking website: red highlights mark differences in the similarity of sensitivity levels. Actual stimuli were shown without highlights. The white boxes symbolize checkboxes that can be clicked in order to share the personal data items.

To vary the choice amount, we configure a small (6 checkboxes) and a large (42) array of options. To further modify the decision complexity independent of the choice amount, we vary the object structure between a homogeneous and heterogenous version (see Fig. 3). Both variables are combined in a 2×2 between-subject experimental design. Table 1 lists all four conditions used in Study 2. Two screenshots of the ACMs in the actual experiment, Figure 10 for Condition 0 and Figure 11 for Condition 3, are provided in Appendix E (translated to English for this presentation).

Condition	Choice amount	Object structure	
0 (control group)	small	homogenous	Decision complexity
1	small	heterogeneous	
2	large	homogenous	
3	large	heterogeneous	

Table 1: Overview of experimental conditions

3.4 Hypotheses

In Sections 2 and 3.1 we have identified the option structure as a moderator of choice complexity. As indicated in Table 1, we expect that an increase in the choice amount amplifies this complexity. The resulting partial order lets us derive four hypotheses:

H1 Participants assigned to Condition 2 report a higher score on the TMC scale than participants assigned to Condition 0.

H2 Participants assigned to Condition 3 report a higher score on the TMC scale than participants assigned to Condition 1.

H3 Participants assigned to Condition 3 report a higher score on the TMC scale than participants assigned to Condition 2.

H4 Participants assigned to Condition 1 report a higher score on the TMC scale than participants assigned to Condition 0.

We refrain from formulating hypotheses on the effects on other measurements, like the PCRT scale, because this is beyond the scope of our decision-making model. Nevertheless, apart from the validation of hypotheses, we strive for an exploration of other, so far unexplained and less salient relations between the proposed measurements, conditions, and the TMC scale.

As we cannot rule out that the types of data recipients might affect the complexity of the disclosure decision, we are interested in the perceived trustworthiness of the subjects presented in the ACM. For instance, if all presented subjects are perceived as very trustworthy, the overall decision complexity might be very low irrespective of choice amount or object structure. A similar argument can be made if all items are perceived as either highly sensitive or not sensitive at all. Therefore, we deem it necessary to control these parameters. Since we cannot measure this information during the actual experiment, we carried out a pre-study (Study 1) to collect empirical data on contextual experience, as well as perceived sensitivity and trustworthiness of the objects and subjects in the ACM, respectively.[3]

[2]See www.linkedin.com, www.xing.com

[3]A comparable methodology is used in [37] and [45]

3.5 Recruitment and Ethical Aspects

Both studies were carried out online. Participants were recruited on a voluntary basis from a German-speaking university campus population, typically via personal invitation by the researchers in lecture halls of a variety of subjects and through word of mouth and social media. No tangible incentives were given and all instruments were compliant with German data protection law. Every participant was reminded to be part of an online study and that all personal data, including survey responses and profile information, will be stored on university-hosted servers.

Regarding ethical aspects, Study 1 is a typical opinion survey that does not involve deception nor touch any overly sensitive topic. Study 2 requires more careful consideration because the experiment was explicitly framed as an usability study of a business networking website, which was claimed to be currently developed by the university. Although the website adopted the corporate design of the university, it intentionally had a salient prototype-like appearance, i. e., most parts of the site were marked as "work in progress" or disabled. The candidates were told that by participating in this study they support the university in improving the usability of the planned service. It was further mentioned that they can share their profile with other participating members including potential employers. To minimize unfulfilled expectations, we did not name any company and further emphasized that the primary function of the network is to get in touch with potential employers and not to serve as a job search tool. In fact, the university already offers comparable services so that the website represents just another communication channel. In line with our expectations, and verified in Study 1, few participants reported to be actively looking for a job. Rather, they were interested in being contacted by local employers in general.

In the debriefing phase of Study 2, we informed the participants that this study was neither a usability study nor a real business networking website. We further mentioned that the purpose of this study was to test different layouts of privacy settings in business networking websites. We also asked our participants for honest comments after the debriefing and we have not received any expressions of disappointment. We further emphasized that all personal data except the survey responses will be deleted immediately.

Note that in Germany, it is primarily the responsibility of the individual researcher to ensure that a planned experiment does not violate research ethics, which are taught at length in many classes. IRBs for this kind of research are not very common. Nevertheless, we sought advice from experienced international researchers whom we met in the context of a summer school. None of them voiced concerns after we presented our empirical approach. Both studies also went through multiple (i. e., at least 20 for Study 2) iterative face-to-face pretests before the actual fieldwork.

4. STUDY 1

The purpose of Study 1 is to calibrate the stimuli for the conditions presented in Section 3.3 and to explore the contextual experience of our population with the scenario. We therefore asked the participants to report if they "never heard" (0), "heard" (1), or "are members" (2) of a business networking website. This *membership indicator* is used as a grouping variable to examine if the answers are invariant to contextual experience. The complete set of items used to measure contextual knowledge and motivations is listed in Table 9 in Appendix C. The main part of the survey asked the participants to imagine the role of a user of a business networking website. We provided additional background on how these networks usually work and what might be potential benefits for subscribers. We asked the participants to rate how comfortable they would be with sharing personal data items on their network profile. We use a 7-point *sensitivity scale* semantically anchored from "very uncomfortable" (1) to "very comfortable" (7) to record the responses. To reduce drop-outs or habituated responses, we divided the personal data items into two groups and distributed them over two survey pages. The order of the personal data items was randomized per subject to attenuate response order effects and to identify inconsistent answers. The *trust scale* asked the participants to rate the trustworthiness of other network members which will appear as subjects (S) in the ACM of the main study. Responses were collected on a 7-point scale, semantically anchored from "not trustworthy at all" (1) to "completely trustworthy" (7). The survey closed with questions on general privacy concerns (adopted from [56, 20]).

4.1 Results

We recruited 60 German-speaking participants and excluded the responses of 6 subjects because they failed to answer several items and revealed inconsistent response patterns. The remaining 54 participants were mainly students (90%, undergraduate and graduate), 25 female and 29 male, with an average age of 24.5 years (range: 18–33).

4.1.1 Contextual Experience

All participants were registered users of a mainstream social networking service and reported to use the service multiples times per week (75% multiple times per day). By contrast, only 29.8% reported to be active members of business networking websites. 25 of the remaining 40 participants (62.5%) have at least heard about such services. A minority of the participants (26.3%) reported to be on active job search, however, 86% answered to be interested in job offers by potential employers. Besides a moderate correlation between age and membership in a business networking website, we could not identify any demographic predictor for context-related items.

4.1.2 Sensitivity and Trustworthiness

In total we ranked 27 different personal data items, listed in Table 7, and 9 subjects, listed in 6 (both Appendix A). All items and subjects have been derived from real world instances of social or business networking websites. A set of test variables (items 25–27, e. g., alcohol consumption) was used to identify participants who did not actively process each option and picked elusive rating scores. The descriptive statistics show that for almost all data items, the full range of rating scores was used. A Shapiro–Wilk test revealed that the rating results did not follow a normal distribution ($p < .001$ for all 27 items). Because of this and the varying group sizes, we computed a series of Kruskal–Wallis one-way analyses. We tried to identify personal data items that are sensitive to different contextual experiences. The tests indicate that there is no significant difference in the medians between the three different levels of the member-

ship indicator. The Kruskal–Wallis test applied to the trustworthiness scale found significant differences within subject "colleagues" ($\chi^2(2, N = 54) = 6.246$, $p < .05$). Pairwise post-hoc comparisons indicate that the mean scores between the groups "never heard" ($M = 3.31$, $SD = 1.888$) and "member" (colleagues: $M = 5.24$, $SD = 1.200$) differ significantly. Therefore, we conclude that individual differences about the trustworthiness of colleagues prevail and therefore decided to remove this subject from the final study. The full set of results including descriptive statistics are reported in Table 6 for the trustworthiness scale and Table 7 for the sensitivity scale (Appendix A).

4.1.3 Clustering

To determine the heterogeneous object structure by grouping personal data items, we computed an inverted distance matrix and fed it to a hierarchical Ward clustering (z-score scaled, Euclidean distance). To determine the homogeneous object structure, a k-means clustering (z-score scaled) was used. We compared the silhouette coefficient and the sum of within-group variances to measure for the quality of the clustering output. Due to the high variance in the scores of the personal data items, the clustering result for the homogeneous groups was expected to be only of moderate quality, which was eventually confirmed by our measurements.

In total we tested solutions with $k = \{1, \ldots, 7\}$ clusters. The k-mean algorithm uses a randomly selected starting point for the clustering process, which also affects the clustering quality. As a remedy, we ran 250 clustering iterations for each k and used the best result as a benchmark. Finally, we used the $k = 5$ solution which is depicted in Figure 9 (Appendix A) to design the final conditions. For the sake of readability, we decided to split group 5 with 13 elements in two groups of 5 and 8 elements, to finally obtain 6 clusters. The accepted solution had a rather weak model quality, but was sufficient to derive clearly distinguishable objects structures, as depicted in Figure 4. Each bar in the figure represents the sums of within-group variances of the four conditions. In both cases the heterogeneous compositions have a stronger variance than either corresponding homogeneous configuration. The final composition of all four cluster results is listed in Table 7 (Appendix A).

4.2 Discussion

Study 1 has helped us to gain valuable insights in how the quantitative experiment should be designed. Both the sensitivity and the trustworthiness scale facilitated the creation of suitable and empirically grounded stimuli for the four conditions in the main experiment. The successful clustering of data collected with randomized item orders demonstrates that participants respond attentively. We interpret this as an indication of generally good data quality.

5. STUDY 2

Study 2 tries to answer our confirmatory and exploratory research questions (Sect. 3.1); the former by testing the hypotheses formulated in Section 3.4, the latter by additional statistical analyses and visualization. The design of Study 2 is more complex than Study 1. Therefore we devote specific subsections to the description of the measurements scales and control variables (Sect. 5.1) and the procedure (Sect. 5.2).

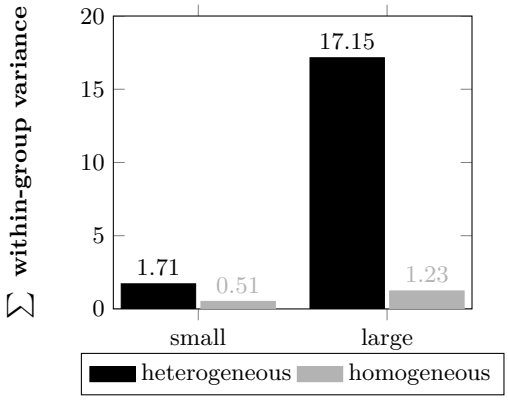


Figure 4: Empirical calibration of stimuli. Sum of the within-group variance for small (13) and large (24) choice conditions, broken down by the object structure. The heterogeneous object structure has higher within-group variance than the homogeneous object structure in both cases. (Study 1)

5.1 Measurement and Controls

Participants are asked to report their experiences, emotions, and opinions in entry and exit surveys. We use two latent dependent variables (DV) and two latent intervening variables (IV), all measured by summing up the responses to at least four indicator questions. This is a common procedure recommended to attenuate response errors on individual questions. To maintain internal consistency, we have eliminated items with a selectivity below 0.30 [21] (see Table 11 in Appendix D). A third intervening variable is collected from a single question with an ordinal scale. In addition to reactive measurements, we collect technical data about the participants' actual behavior as control variables.

5.1.1 Too Much Choice, TMC (DV)

Inspired by the items used in [41] and [50], we measure reported satisfaction, confidence, carefulness, and suitability with regard to the decision process and its outcome. The original question wordings were translated to German and adapted to the context of privacy settings. All responses are collected on 7-point semantically anchored scales. The TMC score is calculated as the sum of six items. It is our main dependent variable that measures immediate reflections of past disclosure decisions. A high TMC score indicates more negative feelings (dissatisfaction, frustration). The aggregated scale ranges from 6 (strong positive) to 42 (strong negative reflection). Post-hoc, this scale had "excellent" reliability as indicated by Cronbach's $\alpha = .922$.

Other published TMC studies either use rating scales of self-reported satisfaction with the decision process and its outcome, or a dichotomous indicator for the deferral of choice, offered by a symbolic no-choice option [16]. We leave deferral options in the privacy domain to future work.

5.1.2 Perceived Comfort, Risk & Trust, PCRT (DV)

Perceived comfort, risk, and trust are relevant factors in the assessment phase of a (disclosure) decision [18, 1]. We repurpose these factors as retrospective measurements to capture potential adverse impacts of general TMC effects. Persons who strongly regret a privacy decision might experience a deterioration of mood and project this negative

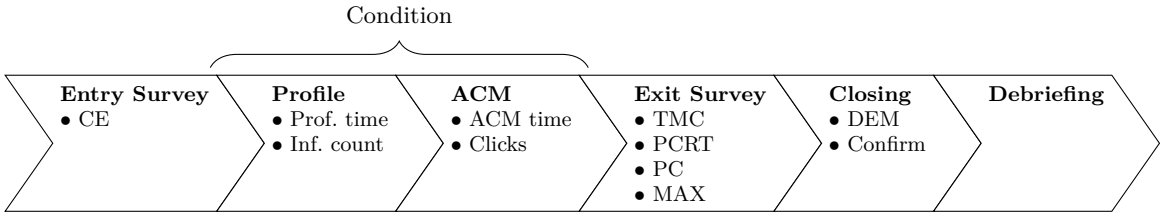


Figure 5: Process of the main experiment. Sequence of phases to be completed by the participants with associated measurement and control instruments. Conditions modify the stimuli presented in the Profile and the ACM phase. (Study 2)

feeling in a broader context than just the specific decision process. The aggregated scale ranges from 4 (strong positive perception of the website) to 20 (strong negative perception of the website). The PCRT scale had "acceptable" reliability as indicated by Cronbach's $\alpha = .693$. The lower reliability compared to the TMC scale is not surprising because the items were put together in an ad-hoc manner rather than by a sophisticated scaling procedure.

5.1.3 Privacy Concerns, PC (IV)

We use a 6-item scale with items adopted from [1, 20, 38] to measure general privacy concerns of the participants. We expect that more privacy-concerned participants produce higher PCRT scores in general. This means they assign a lower trustworthiness to the system and perceive more risks to their privacy. The scale ranges from 6 (strong privacy concerns) to 30 (no privacy concerns at all).

5.1.4 Maximizer, MAX (IV)

Schwartz et al. [52] developed a scale to test a decision makers' inclination of rather pursuing satisficing or optimizing (aka maximizing) strategies. We use 6 items out of the originally proposed 13 to build a scale where a higher value indicates a stronger tendency to maximize. The scale ranges from 6 (strong tendency to satisfice) to 54 (strong tendency to optimize).

5.1.5 Contextual Experience, CE (IV)

We chose to embed our experiment in a business networking website and could not expect that every participant is already familiar with such a service. Against the backdrop that researchers struggle to identify domain-specific expertise as a robust moderator of the TMC effect [49, 50], we decided to use the membership indicator along with other contextual experience questions from Study 1 as controls that allow for further interpretation of the results. Table 9 (Appendix C) reports the full list of questions including filter conditions.

5.1.6 Embedded Controls

The layout of the website mimicked the university's corporate design and the structure of its e-learning and course management platform. Unlike in Study 1, participants in Study 2 had to log-in with a valid university account. During the briefing phase, the participants were addressed with their real name, which we retrieved from the university directory.[4] We did this to reinforce the official character of the website. None of 20 pre-testers doubted that the website is an official university service.

During the experiment, we measured four behavioral control parameters to explain differences in the TMC results: time needed to complete the profile, time spent on the privacy settings, total number of clicks in the ACM, and the number of personal data items entered (excluding obvious nonsense, which we identify on the fly with basic natural language processing). We also queried the size of the browser window and whether scrollbars were displayed in order to control for potential influences of the visual presentation, in particular of the ACM, on the responses.

5.2 Procedure

Figure 5 visualizes the sequential process each participant in Study 2 went through. The second (profile) and third (ACM) phase were introduced by preceding task descriptions, which are reported in Table 8 (Appendix B). We kept the functionality of the website to a minimum in order to avoid distraction from the participants' main tasks. In the profile phase, participants of the small (large) conditions were asked to enter a minimum of 27 (40) types of information in a CV-style profile. Most information types could be entered multiple times (education, work experience). The profile setup was structured as a step-by-step tour through different input forms, asking for different types of information which are commonly used in online social and business networking websites. To overcome potential inhibitions and uncertainties, we provided a descriptive social norm by annotating each input field with a tooltip that provided examples for suitable input values. The layout for small and large conditions differed only by the number of information items that could be entered. The choice structure did not affect the stimulus of the profile phase.

Participants who completed their profile were forwarded to the ACM page, framed as "privacy settings" dialog. The size and structure of each group of personal data items (in rows) as well as the number of subjects (in columns) were determined by the randomly assigned condition (compare Sect. 3.3 for more details).

After adjusting the privacy settings, the participants were forwarded to the exit survey, which contained the items for the TMC, PCRT, PC, and MAX scales. As an icebreaker question, we asked the participants to rate the general usability of the website and the idea of providing such a service by the university. We used reverse-coded items to identify

[4] In line with our privacy policy, the name and account is bound to a session in memory only and not recorded in the research data. For the time of the fieldwork, we use a separate data structure to record hashes of account names for the purpose of detecting and preventing multiple participations by the same account holder.

participants with inconsistent reporting behavior.

In the closing phase, we reminded the participants that their profile will be stored independently of their survey responses and asked them to provide some additional demographic information (DEM). The experiment closed with a confirmation question asking if the responses were truthful enough and whether the record should therefore be included in the analysis or deleted (Confirm).

In the debriefing we informed the participants about the true purpose of the study and that their profile will be deleted for data protection (cf. Table 8 in Appendix B for the wording and Sect. 3.5 for ethical considerations). All participants had the opportunity to give us feedback in an open-ended question.

5.3 Results

We recruited 112 volunteers as participants. Data from thirteen participants were removed because they dropped out or entered obviously false information during steps 1 or 2 (entry survey, profile information), or dropped out in early stages of the exit survey. Recall that information disclosure was on a voluntary basis and no information type was mandatorily requested by the system. As a consequence, some participants entered only small fragments of information. Those participants tended to spend less time on the ACM than other participants in the same condition. To reduce the amount of noise in the data, we decided to remove all records of participants who entered less than 80% of the requested minimum 27 (40) information items. After imposing this restrictions, data from 81 subjects remains in the statistical analysis, with 19–22 cases per condition. The remaining 81 participants were all students (undergraduate and graduate combined), 40 female and 41 male, with an average age of 25.0 years (range: 18–31). Unless otherwise stated, we use ANOVA to test for differences in score means between conditions and Pearson's product moment coefficient to measure correlations between scales.

5.3.1 Main Effect

Figure 6 shows boxplots of the two dependent variables broken down by condition. Participants in the large choice conditions tend to report higher TMC scores than participants in the small choice conditions. Table 2 reports the results of pairwise one-way ANOVAs. The difference in TMC means is statistically significant ($p \leq .05$) in the hypothesized direction between Conditions 0 and 1 and highly significant ($p \leq 0.001$) between Conditions 1 and 3. **Hypotheses H1 and H2 are therefore supported. This contributes to the answer of RQ 1.** We also observe a tendency in line with our expectations for the effect caused by the object structure while the choice amount remains constant. However, the differences in means between Conditions 0 and 1 (small choice amount) as well as between Conditions 2 and 3 (large choice amount) are not statistically significant. **Therefore we reject hypotheses H3 and H4, which are both associated with RQ 2.** The object structure does not seem to raise the TMC score; or not strong enough to distinguish it from noise in our sample.

These results are echoed by the PCRT scale, though statistically less significant (but above the $p \leq .05$ threshold).

We find a positive but low correlation between the TMC and PCRT scale ($r(81) = .168, p > .05$), indicating that both scales measure different negative consequences of choice

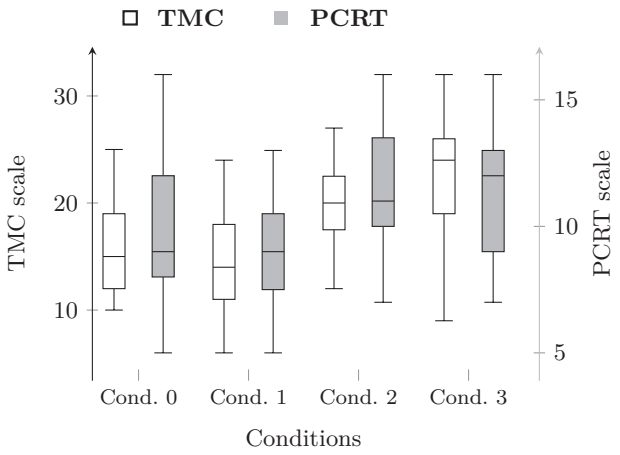


Figure 6: Boxplot of TMC (LHS) and PCRT (RHS) scores by condition. Higher scores for Conditions 2 and 3 indicate that more options negatively affect satisfaction (TMC) and trust (PCRT). (Study 2)

	Comparison	N	TMC	PCRT
H4	Condition 0 < Condition 1	39	.156	.534
H1	Condition 0 < Condition 2	39	4.256*	5.535*
	Condition 0 < Condition 3	42	14.344***	4.498*
	Condition 1 < Condition 2	37	4.389*	10.004**
H2	Condition 1 < Condition 3	42	13.078***	8.561**
H3	Condition 2 < Condition 3	40	3.206	.096

*$p \leq 0.5$, **$p \leq 0.01$,***$p \leq 0.001$

Table 2: Statistical significance tests for differences between conditions in the TMC and PCRT scores. F-values of one-way between-subject ANOVAs, two-sided p-values for robustness. (Study 2)

overload. Moreover, participants with higher privacy concerns (PC) perceive more risks, are less satisfied with the protection of their privacy (PCRT, $r(81) = -.274, p \leq .05$), and are less satisfied with their disclosure decisions (TMC, $r(81) = -.269, p \leq .05$) across all conditions. **This sheds initial light on RQ 3.**

To the best of our knowledge, this is the first empirical evidence that participants who are confronted with larger and more complex personal data disclosure decisions reflect the decision process more negatively in terms of satisfaction, regret, and feeling overwhelmed (i. e., higher scores on the TMC scale) than participants who face a small and less complex disclosure decision. Furthermore, we observe a noticeable negative effect of the large choice condition on the reported trustworthiness of the website and the perceived comfort when using it (higher scores on the PCRT scale). For completeness, Table 3 reports the first two moments for the two dependent variables broken down by condition and individual items.

In summary, our results consistently support a causal influence of choice amount on the evaluation phase of end-user privacy decisions. They also do not rule out the possibility that the choice structure has some impact on the reflection of a decision, but the amount of choice was more decisive in our setup. This is not very surprising, because the manipulation of the structure has more subtle effects and is thus harder to identify in small samples.

DV*	Cond. 0		Cond. 1		Cond. 2		Cond. 3	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
TMC	16.04	4.65	15.36	6.19	19.16	4.87	22.22	5.93
TMC1	2.86	1.15	2.58	1.07	3.21	.79	3.59	1.14
TMC2	2.71	1.15	2.68	1.16	3.32	.89	3.68	.99
TMC3	2.95	.97	2.63	1.26	3.11	.94	3.73	1.24
TMC4	2.52	.75	2.42	1.22	3.37	1.21	3.91	1.15
TMC5	2.67	1.02	2.57	1.22	2.79	.92	3.36	1.18
TMC6	2.33	.97	2.58	1.22	3.37	1.21	3.95	1.39
TMC7	4.10	1.34	3.16	1.39	3.11	1.24	3.36	1.59
PCRT	9.61	2.64	9.05	2.22	11.63	2.77	11.36	2.75
PCRT1	1.43	.51	1.42	.51	1.21	.42	1.50	.97
PCRT2	1.43	.60	1.63	.76	1.47	.77	1.77	.69
PCRT3	1.57	.51	1.42	.61	1.47	.70	1.68	.78
PCRT4	1.62	.59	1.37	.60	2.00	.68	1.73	.63
PCRT5	1.52	.81	1.42	.51	2.16	.96	2.05	.84
PCRT6	1.76	.77	1.74	.73	2.21	.86	2.14	.83

* Wording provided in Table 11

Table 3: Means and standard deviations of the complete TMC and PCRT item pool (no items excluded) broken down by condition. (Study 2)

5.3.2 Individual Differences in Decision Strategies

Following [52], we compute a median split of the MAX scores to distinguish between statisficers and optimizers. We did this to ascertain that both characteristics are equally distributed over all conditions (cf. Table 4). We use the full scale score to test if the tendency to optimize affects the individual TMC score. We find a moderate correlation between the MAX and TMC scores in both small choice amount conditions (significant only for Condition 1, cf. Figure 7), but not for the large choice amount conditions. We conjecture that personal traits and habits influence the overall rating score more in simple decisions than in cases which require more systematic processing for the sheer size of the decision space. **This adds to a partial answer of RQ 3**, but more research is needed to fully understand the underlying mechanism.

5.3.3 Demographics and Contextual Experience

We find no significant differences in the demographics between conditions (cf. Table 4). This is reassuring because demographic attributes apparently do not cause differences in drop-out rates or reported scores. The distribution of the membership indicator over conditions groups is also reported in Table 4. We cannot find a sign of statistical dependence between the TMC score and the membership indicator, neither in total nor within conditions. The same holds for all other reported demographics. This is in line with our findings in Study 1, where neither contextual experience nor age or gender had a significant influence on other ratings.

5.3.4 Embedded Controls

We estimate a linear multiple regression model per group to capture the relation between the control mechanisms and the TMC score. Analyses across groups are out of the scope of this paper because groups are not directly comparable for some controls. The following parameters are included as predictors: the time spent on the privacy settings page (*privacy_time*), the total number of clicks in the ACM (*clicks*), the number of data sharing permissions (*total_shared*), and the number of personal data items entered in the profile (*answered_fields*). We are aware of potential multicollinearity issues. In particular, *clicks* and *total_shared* are closely associated. However, in all cases the variance inflation factor varied in the permissible range between 1 and 3 [40].

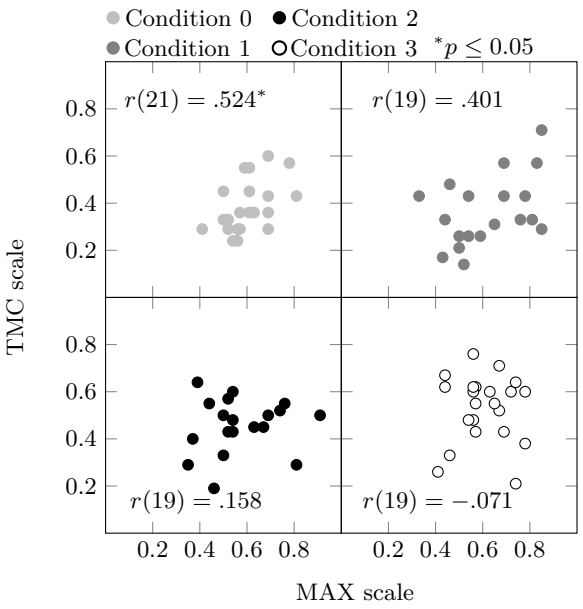


Figure 7: Scatterplots and Pearson product-moment correlation of normalized TMC and MAX scores. In conditions 0, 1, and 2 maximizers tend to be less satisfied with their choices made. (Study 2)

As indicated by the regression results shown in Table 5, the most stable predictor across all conditions is *clicks* followed by *total_shared* and the number of personal data items entered in the profile. All of them are positively associated with the TMC score. The time spent on the privacy settings page appears to be the weakest predictor and is correlated even negatively in one case. However, very few of the predictors differ significantly from zero. This may be partly due to the relatively high number of predictors compared to the sample size. In general, the models for the large condition groups explain a larger share of the variance in the TMC score. This corroborates the conjecture in Sect. 5.3.2 that the TMC score is dominated by the condition if the decision space is large, thereby displacing other factors that tend to have a smaller and more heterogeneous influence.

A series of product-moment correlations computed between the overall TMC scores and the four predictors identifies the number of clicks as strongest covariate ($r(81) = .614, p \leq .001$, see Figure 8 for a visualization).

5.4 Discussion

The results of Study 2 can be divided in results related to the experimental factors, which permit a causal interpretation, and results related to individual traits, which are self-reported and therefore prone to endogeneity issues.

The experiment revealed that a larger number of data sharing options causes significantly more negative emotional reactions in the evaluation phase of a decision process, as reported on established items of the TMC scale. The results confirm the hypothesized negative impact of choice proliferation on satisfaction, the experience of regret, and feel-

Condition	N	Gender (%)		Age		Decision strategy (%)		Membership indicator (%)		
		Female	Male	Mean	Max	Optimizer	Satisficer	Never Heard	Heard	Member
Condition 0	21	47.6	52.4	25.48	29	47.62	52.38	9.5	66.7	23.8
Condition 1	19	57.9	42.1	25.32	30	47.37	52.63	5.3	57.9	36.8
Condition 2	19	52.6	47.4	24.37	29	36.84	63.16	5.3	68.4	26.3
Condition 3	22	40.9	59.1	24.86	31	45.46	54.54	13.8	54.5	31.7
Total	81	49.8	50.2	25.01	31	44.32	55.68	8.5	61.8	29.7

Table 4: Demographics, decision strategy, and membership indicator by condition. (Study 2)

Regression	clicks		total_shared		privacy_time		answered_fields		Constant		Model	
	Coef.	t	Coef.	t	Coef.	t	Coef.	t	Coef.	t	R^2 adj.	F
Condition 0	.953	1.056	3.323	1.130	−.144	−.532	.093	.191	4.254	.332	−.049	.766
Condition 1	2.017	1.539	3.113	1.098	.111	.457	.682	.851	−18.550	−.719	0.58	1.276
Condition 2	1.252*	2.249	.246	.453	.052	.284	.110	.354	−1.994	−.272	.487**	5.276
Condition 3	1.490**	2.992	.021	.019	.093	.614	.345	1.067	−10.589	−.796	.465**	5.569

*$p \leq .05$, **$p \leq 0.01$,***$p \leq 0.001$

Table 5: Results of linear multiple regressions, one per condition. Dependent variable: TMC score. The number of clicks in the ACM and the total number of data sharing permissions are the strongest positive predictors of the TMC score. (Study 2)

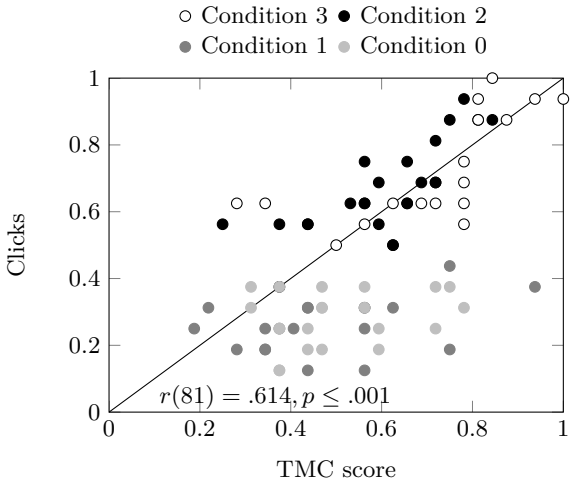


Figure 8: Scatterplot and Pearson product-moment correlation of normalized TMC scores and number of clicks. More clicks are associated with higher dissatisfaction with the choices made. (Study 2)

ings of being overwhelmed. The significantly higher PCRT scores in the conditions with large choice amount indicate that having more choice can also bias the perceived comfort, risk, and trustworthiness of the choice context (i. e. the website). The reported TMC and PCRT scores correlated positively, yet not significantly. This supports our assumption that the negative reflection of the decision, captured by the TMC score, spills over to a negative perception of the overall scenario. It remains the task of future research to investigate more into the causal links between the items perceived comfort, risk, and trustworthiness and the emotional reflection of the decision. To this end, it would be desirable to measure these latent factors with separate item batteries established in psychology. A combined scale, like our PCRT scale, is just a first exploratory step to test for a general relation with choice proliferation.

The analysis on individual differences in the TMC scores revealed mixed findings. Schwartz et al. [52] argue that people who tend to be optimizers suffer more from choice proliferation than satisficers. This theory is only partly confirmed by our results. Participants in the small conditions who reported comparable high TMC scores were more often classified as optimizers. On the other hand, this trait had almost no influence on the TMC scores in the large conditions. We encountered a similar pattern for the controls *clicks*, *total_shared* and *answered_fields*. All three predictors have a stronger influence on individual TMC scores if the choice amount was small. The sample in this work is too small to identify interactions between self-reported traits or control variables and the effect size of the experimental factors.

The hypothesized impact of the object structure on the TMC scale could not be confirmed in our experiment. Recall that our hypotheses are based on the assumption that people perceive choice as less difficult if the option structure is more similar. However, other researchers state that "introducing a small difference in an otherwise identical attribute can increase the perceived similarity of choice alternatives" [31]. This *similarity effect* might have suppressed the predicted effect of the heterogeneous object structure on choice complexity. Follow-up studies should therefore control for this factor, for example by using a quantitative procedure (clustering) to derive different option structures that are presented to human subjects for a rating of the perceived similarity.

6. LIMITATIONS

Although the specification of a decision model and the pre-study helped us to optimize our stimuli, we had to keep the scenario as realistic as possible to ensure external validity. This required compromises by relaxing controls in the experimental design.

For example, the participants were neither compelled to enter all information nor bound to a strict procedure. Hence, each participant experienced the study in a slightly differ-

ent way. As a remedy, we removed records of participants who provided less than 80% of the requested information types. This leads to a more homogenous sample, but might have introduced a bias against more privacy-aware persons, as confirmed by inspecting the PC scores of the excluded records. In general, the well-known limitations of small convenience samples in a student population apply.

As a side effect of manipulating the choice amount, each condition came with visible changes in the profile and ACM. We considered manipulating the ACM only, but were afraid of confusing people by asking for data items that do not appear in the privacy settings. This may elicit feelings of limited control. Another difficulty is that extending the ACM involves adding objects (rows) and subjects (columns). This changes the object and subject structure of the entire decision. Such structural differences are hard to control and may be confounded with the effect of choice amount.

In the large choice conditions, the ACM dominated the layout and emphasized the complexity on a visual level. The same holds for all TMC studies, but in other contexts, the presence of a large choice amount often induces positive feelings at the first glance. We are concerned that this might not hold for our ACM matrix and the privacy domain in general. The TMC effect might have been stronger driven by the visual interpretation of the choice than in common consumer experiments. Although we stressed the benefits of sharing the data ("interesting potential employers might contact you"), this positive consequences may only materialize in the future and are therefore less salient. As a result, the participants might have perceived the task more as a burden instead of having the possibility of choosing among various different options, which all appear very attractive to them at present.

Another difference to conventional TMC studies is that psychology and marketing researchers present 1-out-of-n decisions. Strictly speaking, our privacy settings asked for n binary decisions, which are not necessarily independent. The overall decision complexity may grow disproportionally if the decision maker tries to strive for some sort of consistency. This may amplify the TMC effect in our setting.

Moreover, potential priming and response order effects of the exit survey phase cannot be excluded. In particular the questions asking for trust and risks, placed before and after the TMC question block, might have biased the participants' interpretation of the study.

Finally, the field of TMC research struggles to reproduce many published results and is still seeking for a comprehensive psychological understanding of the TMC effect in general [50]. Some authors even question the existence of a TMC effect in general and point out the lack of robustness against differences in cultures, context, an individual traits [50]. Therefore, our initial evidence in the privacy domain, obtained with a small and homogeneous sample, should be interpreted with caution and not used for policy advice unless the effect is replicated with independent data.

7. SUMMARY AND CONCLUSION

Our study provides initial empirical evidence of negative psychological effects triggered by the proliferation of choice in a privacy context. We use elements of decision field theory, consumer psychology and findings of TMC research in order to devise a model that illustrates selected aspects of a disclosure decision. We report the results of a comprehensive empirical study, a university-hosted business networking website, carried out to test our hypotheses with a quantitative 2×2 experiment. An adapted access control matrix served to simulate the disclosure decision with varying amount and structure of elements, depending on the randomly assigned condition. A pool of established items was used to derive four reliable scales: Too Much Choice (TMC), Perceived Comfort, Risk, and Trustworthiness (PCRT), Privacy Concerns (PC), and Maximizer (MAX).

We find that participants assigned to a large choice condition report to be less satisfied with their choices made, experience more regret, and are more overwhelmed by the decision process. Despite some limitations, we can successfully demonstrate that the number of privacy options presented to a user affects the (short-term) emotional reflection of the decision in the evaluation phase of a decision-making process. Additional exploratory analyses suggest that also the perceived comfort, risk, and trustworthiness of the decision context can be negatively affected by choice proliferation.

Applying this lens to privacy research breaks new ground. While research in psychology discerns the evaluation phase as an important phase of decision-making, privacy research so far seems to be focused on the assessment phase. Researchers try to understand why a decision maker assigns a higher value to the prospect "disclose" than "conceal". Also many interdisciplinary studies contribute to this research by incorporating psychological elements like trust, perceived risks and other concepts from behavioral economics. But even these psychological models are mostly applied to better understand the outcome and not the emotional consequences of decision making. Although there are a few studies which investigate why users regret the outcome of a disclosure decision, they do not capture the actual emotional reflection of the decision processes [62, 42] or use ad-hoc rather than established scales to measure the dependent variable [25]. This work demonstrates that in particular the investigation of a variety of emotional and psychological factors can provide new and valuable insights into end-users privacy decisions.

This work also contributes to the emerging literature which questions the policy trend of putting consumers in charge of controlling the dissemination of their personal data. Against the backdrop of a vastly growing data industry, this criticism appears counter-productive at first sight. However, consumers' privacy decisions are prone to manipulation by subtle changes of the decision context and the choice architecture. A concern commonly raised by privacy advocates is the possibility of strategic abuse of privacy choice architecture by data-intensive industries towards nudging consumers into disclosing personal information above a socially optimal level [19, 5, 4]. Our results suggest that if this implies that more and more disclosure and sharing decision are delegated to the consumer, this not only affects the users' sharing attitudes (identified in [12]) and unnecessarily consumes cognitive resources (as in [9]), but also has measurable emotional consequences in the short run. (We cannot say anything about longer-term effects.) This reinforces the recommendation to designers of privacy panels to not only focus on the layout and composition of privacy settings, but also follow choice minimizing principles and scrutinize the necessity of each additional option. It also reinforces ideas of automating end-user privacy decisions either by safe defaults or with appropriate standards and tool support.

8. ACKNOWLEDGEMENTS

We thank the volunteers who participated in our studies, the attendants of the 2013 Late Autumn School on Online Communication and Online Trust and the anonymous reviewers for advice; and our shepherd Simson Garfinkel for his help in making this paper digestible for SOUPS.

9. REFERENCES

[1] M. S. Ackerman, L. F. Cranor, and J. Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 1–8. ACM, 1999.

[2] A. Acquisti. Nudging privacy. *IEEE Security & Privacy*, 7(6):0082–85, 2009.

[3] A. Acquisti and J. Grossklags. What can behavioral economics teach us about privacy. In *Digital Privacy: Theory, Technologies, and Practices*, page 329. Auerbach Publications, 2007.

[4] A. Acquisti, L. John, and G. Loewenstein. What is privacy worth. In *21st Workshop on Information Systems and Economics (WISE)*, pages 14–15, 2009.

[5] I. Adjerid, A. Acquisti, L. Brandimarte, and G. Loewenstein. Sleights of privacy: framing, disclosures, and the limits of transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*, page 9. ACM, 2013.

[6] K. J. Arrow. *Social Choice and Individual Values*, volume 12. Yale university press, 2012.

[7] R. Balebako, P. G. Leon, H. Almuhimedi, P. G. Kelley, J. Mugan, A. Acquisti, L. F. Cranor, and N. Sadeh. Nudging users towards privacy on mobile devices. In *CHI 2011 Workshop Article*, 2011.

[8] J. R. Bettman, M. F. Luce, and J. W. Payne. Constructive consumer choice processes. *Journal of Consumer Research*, 25(3):187–217, 1998.

[9] R. Böhme and J. Grossklags. The security cost of cheap user interaction. In *Proceedings of the New Security Paradigms Workshop (NSPW)*, pages 67–82. ACM, 2011.

[10] R. Böhme and S. Köpsell. Trained to accept?: A field experiment on consent dialogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, pages 2403–2406. ACM, 2010.

[11] J. Bouckaert and H. Degryse. Opt in versus opt out: A free-entry analysis of privacy policies. In *Workshop of Economics and Information Security (WEIS)*, Cambridge, UK, 2006.

[12] L. Brandimarte, A. Acquisti, and G. Loewenstein. Misplaced confidences privacy and the control paradox. *Social Psychological and Personality Science*, 4(3):340–347, 2013.

[13] J. R. Busemeyer and J. T. Townsend. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3):432, 1993.

[14] F. H. Cate. The limits of notice and choice. *IEEE Security & Privacy*, 8(2):59–62, 2010.

[15] D. I. Cordova and M. R. Lepper. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4):715, 1996.

[16] R. Dhar. Consumer preference for a no-choice option. *Journal of Consumer Research*, 24(2):215–231, 1997.

[17] R. Dhar, S. M. Nowlis, and S. J. Sherman. Trying hard or hardly trying: An analysis of context effects in choice. *Journal of Consumer Psychology*, 9(4):189–200, 2000.

[18] C. Dwyer, S. R. Hiltz, and K. Passerini. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. In *Proceedings of AMCIS*, page 339, 2007.

[19] S. Egelman, A. Felt, and D. Wagner. Choice architecture and smartphone privacy: There's a price for that. In R. Böhme, editor, *The Economics of Information Security and Privacy*, pages 211–236. Springer Berlin Heidelberg, 2013.

[20] European Commission. The gallup organization: Data protection in the european union citizens' perceptions. http://ec.europa.eu/public_opinion/flash/fl_225_en.pdf, 2008. [Online; accessed 27-February-2014].

[21] B. Everitt and A. Skrondal. *The Cambridge Dictionary of Statistics*, volume 3rd Edition. Cambridge University Press, Cambridge, 2006.

[22] B. Fasolo, R. Hertwig, M. Huber, and M. Ludwig. Size, entropy, and density: What is the difference that makes the difference between small and large real-world assortments? *Psychology & Marketing*, 26(3):254–279, 2009.

[23] M. Friedman and R. Friedman. *Free to Choose: A Personal Statement*. Houghton Mifflin Harcourt, 1990.

[24] G. Gigerenzer and P. M. Todd. Fast and frugal heuristics: The adaptive toolbox. In *Simple Heuristics that Make us Smart. Evolution and Cognition.*, pages 3–34. Oxford University Press, New York, 1999.

[25] N. S. Good, J. Grossklags, D. K. Mulligan, and J. A. Konstan. Noticing notice: a large-scale experiment on the timing of software license agreements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*, pages 607–616. ACM, 2007.

[26] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, pages 71–80. ACM, 2005.

[27] G. A. Haynes. Testing the boundaries of the choice overload phenomenon: The effect of number of options and time pressure on decision difficulty and satisfaction. *Psychology and Marketing*, 26(3):204–212, 2009.

[28] H. Heckhausen and P. M. Gollwitzer. Thought contents and cognitive functioning in motivational versus volitional states of mind. *Motivation and Emotion*, 11(2):101–120, 1987.

[29] S. S. Iyengar and M. R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6):995, 2000.

[30] R. K. Jessup, E. S. Veinott, P. M. Todd, and J. R. Busemeyer. Leaving the store empty-handed: Testing explanations for the too-much-choice effect using decision field theory. *Psychology & Marketing*, 26(3):299–320, 2009.

[31] J. Kim, N. Novemsky, and R. Dhar. Adding small

differences can increase similarity and choice. *Psychological Science*, 24(2):225–229, 2013.

[32] B. P. Knijnenburg, A. Kobsa, and H. Jin. Preference-based location sharing: are more privacy options really better? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pages 2667–2676. ACM, 2013.

[33] J. Lafferty. Default privacy settings for new teens' facebook accounts set at friends-only, adds public option. http://www.insidefacebook.com, 2013. [Online; accessed 01-March-2014].

[34] B. W. Lampson. Protection. *ACM SIGOPS Operating Systems Review*, 8(1):18–24, 1974.

[35] H. M. Lefcourt. The function of the illusions of control and freedom. *American Psychologist*, 28(5):417, 1973.

[36] G. Loewenstein and J. S. Lerner. The role of affect in decision making. *Handbook of Affective Science*, 619(642):3, 2003.

[37] M. Malheiros, S. Preibusch, and M. Sasse. "fairly truthful": The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure. In M. Huth, N. Asokan, S. Čapkun, I. Flechais, and L. Coles-Kemp, editors, *Trust and Trustworthy Computing*, volume 7904 of *Lecture Notes in Computer Science*, pages 250–266. Springer Berlin Heidelberg, 2013.

[38] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[39] J. Nickel and H. Schaumburg. Electronic privacy, trust and self-disclosure in e-recruitment. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*, pages 1231–1234. ACM, 2004.

[40] R. M. O'brien. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690, 2007.

[41] A. Oulasvirta, J. P. Hukkinen, and B. Schwartz. When more is less: the paradox of choice in search engine use. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 516–523. ACM, 2009.

[42] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee. Reasons, rewards, regrets: privacy considerations in location sharing as an interactive practice. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*, page 5. ACM, 2012.

[43] J. W. Payne, J. R. Bettman, and E. J. Johnson. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534, 1988.

[44] S. Pötzsch, P. Wolkerstorfer, and C. Graf. Privacy-awareness information for web forums: Results from an empirical study. In E. T. Hvannberg, M. K. Lárusdóttir, A. Blandford, and J. Gulliksen, editors, *Proceedings of the 6th Nordic Conference on Human-Computer Interaction (NordiCHI)*, pages 363–372. ACM, 2010.

[45] S. Preibusch, K. Krol, and A. R. Beresford. The privacy economics of voluntary over-disclosure in web forms. In R. Böhme, editor, *The Economics of Information Security and Privacy*, pages 183–209. Springer Berlin Heidelberg, 2013.

[46] F. Rothbaum, J. R. Weisz, and S. S. Snyder. Changing the world and changing the self: A two-process model of perceived control. *Journal of Personality and Social Psychology*, 42(1):5, 1982.

[47] J. B. Rotter. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1):1, 1966.

[48] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68, 2000.

[49] B. Scheibehenne, R. Greifeneder, and P. M. Todd. What moderates the too-much-choice effect? *Psychology & Marketing*, 26(3):229–253, 2009.

[50] B. Scheibehenne, R. Greifeneder, and P. M. Todd. Can there ever be too many options? a meta-analytic review of choice overload. *Journal of Consumer Research*, 37(3):409–425, 2010.

[51] B. Schwartz. *The Paradox of Choice*. HarperCollins, 2009.

[52] B. Schwartz, A. Ward, J. Monterosso, S. Lyubomirsky, K. White, and D. R. Lehman. Maximizing versus satisficing: happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5):1178, 2002.

[53] M. E. Seligman. *Helplessness: On depression, development, and death.* WH Freeman/Times Books/Henry Holt & Co, 1975.

[54] C. E. Shannon. A mathematical theory of communications. *Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[55] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129, 1956.

[56] H. J. Smith, S. J. Milberg, and S. J. Burke. Information privacy: measuring individuals' concerns about organizational practices. *MIS quarterly*, pages 167–196, 1996.

[57] D. J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–564, 2006.

[58] K. Tang, J. Hong, and D. Siewiorek. The implications of offering more disclosure choices for social location sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, pages 391–394. ACM, 2012.

[59] R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, 2008.

[60] A. Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281, 1972.

[61] A. Tversky and S. Sattath. Preference trees. *Psychological Review*, 86(6):542, 1979.

[62] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. I regretted the minute i pressed share: A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS '11)*, page 10. ACM, 2011.

[63] A. F. Westin. *Privacy and Freedom*. Atheneum, New York, 1967.

APPENDIX

A. EMPIRICAL RESULTS, STUDY 1

	Descriptive Statistics[a]		Kruskal-Wallis (omnibus)[b]	
Type of network group	**M**	**SD**	$\chi^2(2, N = 54)$	**p**
Friends	5.60	1.405	3.038	.219
Fellow students	4.46	1.342	4.514	.105
Colleagues	4.35**	1.739	5.286	.008
Post-Hoc				
Heared, Member			2.396	.122
Never heared, Member			7.530	.006
Heared, Never heared			2.273	.132
Favorite employers	4.04	1.822	2.365	.307
Employers of a selected industry	3.92	1.702	.890	.641
Employment agency	3.46	1.756	2.674	.263
University employees	3.11	1.354	5.299	.071
All employees	3.02	1.596	.297	.862
All network members	1.81	1.150	2.210	.331

[a] Aggregated, Trustworthiness scale (1) Not trustworthy at all , (7) Completely trustworthy
[b] Grouping variable: membership indicator $^{**}p \leq 0.01$

Table 6: Check for invariance of the median (rank), grouped by membership indicator. $N = 54$**. (Study 1)**

		Cluster allocation		Sensitivity[a]		Kruskal-Wallis (omnibus)[b]	
#	**Information Entity**	Large Hom(Het)[c]	Small Hom(Het)	**M**	**SD**	$\chi^2(2, N = 54)$	**p**
1	Given name	1 (6)	1 (1)	5.72	1.45	4.219	.121
2	Family name	1 (4)	1 (1)	4.96	1.85	2.042	.360
3	Mobile number	6 (3)	-	1.79	1.46	4.424	.109
4	Age	1 (3)	1 (2)	5.62	1.39	.442	.802
5	Favorite food	5 (5)	-	3.00	2.17	.398	.819
6	Favorite TV show	5 (4)	-	2.30	1.78	.276	.871
7	Relationship status	4 (1)	-	1.98	1.52	.392	.822
8	Political interests	4 (3)	-	2.52	1.65	1.735	.420
9	Practiced sports	1 (3)	1 (2)	4.15	1.75	1.830	.400
10	Instant messenger number	6 (6)	-	2.35	1.62	2.507	.483
11	Gender	1 (2)	1 (1)	6.19	1.33	.168	.919
12	Favorite computer game	4 (2)	-	1.81	1.52	1.069	.586
13	Technical expertise	2 (1)	2 (2)	5.89	1.13	.892	.640
14	Job experience	2 (1)	2 (1)	4.81	1.84	5.722	.057
15	Current university grade point average	3 (3)	-	3.54	2.01	2.109	.348
16	Transcript of records (education)	3 (5)	-	3.48	1.80	3.430	.180
17	Education	2 (5)	2 (2)	5.31	1.60	1.969	.374
18	Social skills	2 (4)	2 (2)	5.75	1.26	4.167	.124
19	Language skills	2 (3)	2 (1)	5.89	1.21	.285	.252
20	Attended lectures (university)	2 (4)	2 (1)	4.99	1.70	1.315	.518
21	Selected university records	3 (2)	-	4.28	1.93	.174	.917
22	Received awards	2 (6)	2 (1)	5.24	1.62	4.725	.094
23	Topics of thesis	2 (3)	2 (2)	5.14	1.83	3.743	.154
24	Desired salary	3 (4)	-	3.26	1.68	1.678	.432
25	Medical records[d]	-	-	1.52	1.24	.889	.641
26	Favorite alcoholic drink[d]	-	-	1.40	1.03	2.149	.341
27	Frequency of alcohol consumption[d]	-	-	1.28	.818	.694	.707

[a] Aggregated [b] Grouping variable: membership indicator [c] Hom = Homogeneous, Het = Heterogeneous [d] Removed test items.

Table 7: Clustered personal data items (column 1-2). Cluster results (column 3-4). Descriptive statistics for sensitivity scale (1) "Very uncomfortable" , (7) "Very comfortable" (column 5-6). Check for invariance of the median (rank), grouped by membership indicator (column 7-8).$N = 54$**. (Study 1)**

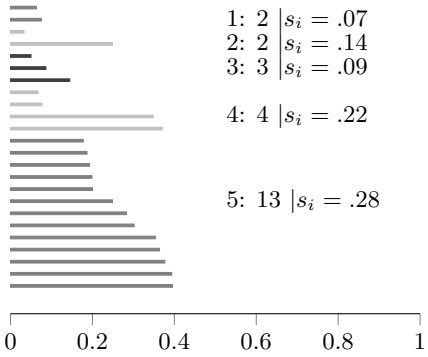


Figure 9: Average silhouette widths (x-axis: s_i) of 24 clustered information personal data items (y-axis). Best result of k-mean clustering with $k = 5$ and 250 iterations is displayed. Total average silhouette width $s_i/n = .22$. (Study 1)

B. TASK DESCRIPTION & DEBRIEFING, STUDY 2

Briefing prior to completion of CV-style profile[a)]
You can configure your user profile on the following pages. Try to keep distractions and interruptions up to a minimum while proceeding with this step! You have the chance to adjust your privacy settings after you have finished this step. Please try to answer all questions honestly and conscientiously. If you do not have particular information at hand, feel free to enter preliminary information which you have in mind. The information can be corrected afterwards. Please note that all answers and information are provided on a voluntary basis.
Briefing prior to managing privacy settings via the ACM
You can manage your privacy settings on the following page. The setting allows you to decide with whom you want to share all or parts of the information you entered. Your privacy is important to us! Take your time to find the privacy settings that you favor the most. To allow a person/group to view your information, you must tick the corresponding checkbox.
Debriefing
For privacy reasons, we will delete your entered profile information after you closed this site. The purpose of this study was to test different privacy settings in social/business networking websites. Please note that this study is not connected with any official university student service. If you have any questions concerning your privacy or are interested in the results of the study, you can contact us by email or leave your e-mail address here: [textfield].

[a)] Original text in German and screenshots available upon request.

Table 8: Task description and debriefing. (Study 2)

C. CONTEXTUAL BACKGROUND AND MOTIVATION, STUDY 1 & 2

Item ID	Wording[a)]
1	Are you actively searching for a new or another employer? ☐ Yes, ☐ No
2	Would you be interested if an employer approached you with a job offer? ☐ Yes, ☐ No
3	Are you a member of a business networking website like Xing or LinkedIn? ☐ Yes, ☐ No
3.1 if (3==no)	Have you ever heard of business networking websites like Xing or LinkedIn? ☐ Yes, ☐ No
3.2 if (3==yes)	Since when are you member of a business networking website? ☐ one month, ☐ one year, ☐ two years, ☐ three years or longer
3.3 if (3==yes)	How often are you using the business networking website ☐ multiple times per day, ☐ once a day, ☐ multiple times per week but less than once a day, ☐ once a week, ☐ less than once a week, ☐ more than once a month but less than once a week, ☐ less than once a week, ☐ I am a member but never used the site
4[b)]	Are you a member of an online social network like Facebook? ☐ Yes, ☐ No
4.1 if (4==no)	Have you ever heard of online social networks like Facebook? ☐ Yes, ☐ No
4.2 if (4==yes)	Since when are you a member of an online social network? ☐ one month, ☐ one year, ☐ two years, ☐ three years or longer
4.3 if (4==yes)	How often are you using the online social network? ☐ multiple times per day, ☐ once a day, ☐ multiple times per week but less than once a day, ☐ once a week, ☐ less than once a week, ☐ more than once a month but less than once a week, ☐ less than once a week, ☐ I am a member but never used the site

a) Original questions in German; wording and screenshots are available upon request.
b) Not used in Study 2.

Table 9: Contextual background and motivation of participants. (Study 1 & 2)

D. EXIT SURVEY AND RESULTS, STUDY 2

Item *	Anchors			
	Min (left)		Max (right)	
TMC1	Very unsatisfied	(1)	Very satisfied	(7)
TMC2	Very hard	(1)	Very easy	(7)
TMC3	No regret	(1)	Strong regret	(7)
TMC4	Not overwhelmed	(1)	Completely overwhelmed	(7)
TMC5	Not frustrating	(1)	Completely frustrating	(7)
TMC7	Completely insufficient	(1)	Completely sufficient	(7)
TMC6	Very unlikely	(1)	Very likely	(7)
PCRT1–6	Completely agree	(1)	Completely disagree	(5)
PC1–7	Completely agree	(1)	Completely disagree	(5)
MAX1–6	Completely disagree	(1)	Completely agree	(7)

* Wording and statistics are provided in Table 11. Original anchors in German.

Table 10: Left and right semantic anchor of all rating scales for the items in Table 11. (Study 2)

Scale	Item	Wording[c)]	Mean	SD	Item total correlation[a)]
TMC	TMC1 (RC)	How satisfied are you with the privacy settings you selected?	3.07	1.104	.803
$N = 81$	TMC2 (RC)	How easy was the selection of the appropriate privacy settings?	3.11	1.118	.806
Mean = 18.30	TMC3	Do you regret the privacy settings made and if so, how much?	3.12	1.166	.832
Reliability[b)]: .922	TMC4	To which extent have you been overwhelmed by choosing the appropriate privacy settings?	3.07	1.243	.707
Range: $[6, 42]$	TMC5	How frustrating was the selection of the correct privacy settings for you?	2.84	1.101	.702
Scale: 7-point semantic	TMC6	Would you choose to correct the privacy settings if this option was available?	3.07	1.358	.724
	~~TMC7~~	Did you think the available privacy settings are sufficient or insufficient?	3.44	1.432	.249
PCRT	~~PCRT1~~	This website appears to be very trustworthy.	1.40	.540	.275
$N = 81$	PCRT2	The risk of entering personal data into this website is low.	1.58	.705	.452
Mean = 6.59	PCRT3	I have the feeling that the personal data I entered are sufficiently protected.	1.54	.672	.449
Reliability: .693	PCRT4	I would not mind using this website again.	1.68	1.790	.591
Range: $[4, 20]$	PCRT5	I felt comfortable using this site.	1.79	.847	.489
Scale: 5-point Likert-type	~~PCRT6~~	I was confident at any time that I have full control over the use of my personal data.	1.96	.813	.265
PC	PC1	I am annoyed by companies who ask for my personal data.	1.96	.749	.553
$N = 81$	PC2	I take care not to give my personal data to Internet companies.	2.00	.758	.770
Mean = 11.81	PC3	The use of personal data should always be bound to a specific purpose.	1.96	.732	.704
Reliability: .769	PC4	I am more concerned about the disclosure of my personal data on the Internet than most other people.	2.59	.959	.442
Range :$[6, 30]$	PC5	Companies are collecting too much of my personal data.	1.60	.736	.341
Scale: 5-point Likert-type	PC6	Companies should invest more to prevent misuse of my personal data.	1.69	.736	.314
	~~PC7~~	Intelligence agencies like the NSA are collecting too much of my personal data.	1.27	.448	.161
MAX	MAX1	No matter how satisfied I am with my job, it is only good for me to watch out for better opportunities.	4.99	1.677	.698
$N = 81$	MAX2	Whenever I am faced with a choice, I try to imagine what all the other possibilities are, even the ones that are not present at the moment.	5.57	1.650	.544
Mean = 32.37	MAX3	When I watch TV, I channel surf, often scanning through the available options even while attempting to watch one program.	5.40	1.794	.345
Reliability: .757	MAX4	I often find it difficult to shop for a gift for a friend.	5.56	1.817	.476
Range: $[6, 54]$	MAX5	I am a big fan of lists that attempt to rank things (the best movies, the best singers, the best athletes, the best novels, etc.).	5.43	1.774	.451
Scale: 9-point semantic	MAX6	No matter what I do, I have the highest standards for myself.	5.43	1.774	.496

[a)] Measured for complete scale, i. e, prior to item exclusions. [b)] Measured with Cronbach's α.
[c)] Original questions in German; wording and screenshots are available upon request.
RC = Reverse coded

Table 11: Item pool of the TMC, PCRT, PC, and MAX scales (Study 2). Items with total correlation $\leq .3$ were excluded (crossed out). The summary statistics in the first column apply to the final aggregated scales.

E. SCREENSHOTS

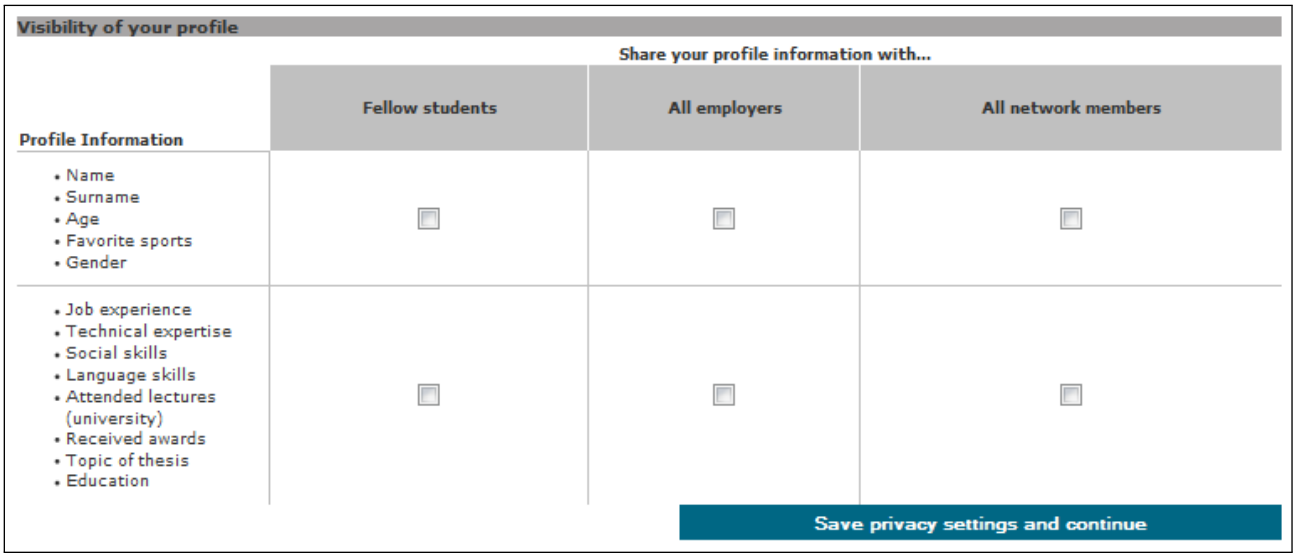


Figure 10: ACM layout for Condition 0: small choice amount and homogenous object structure. (Study 2)

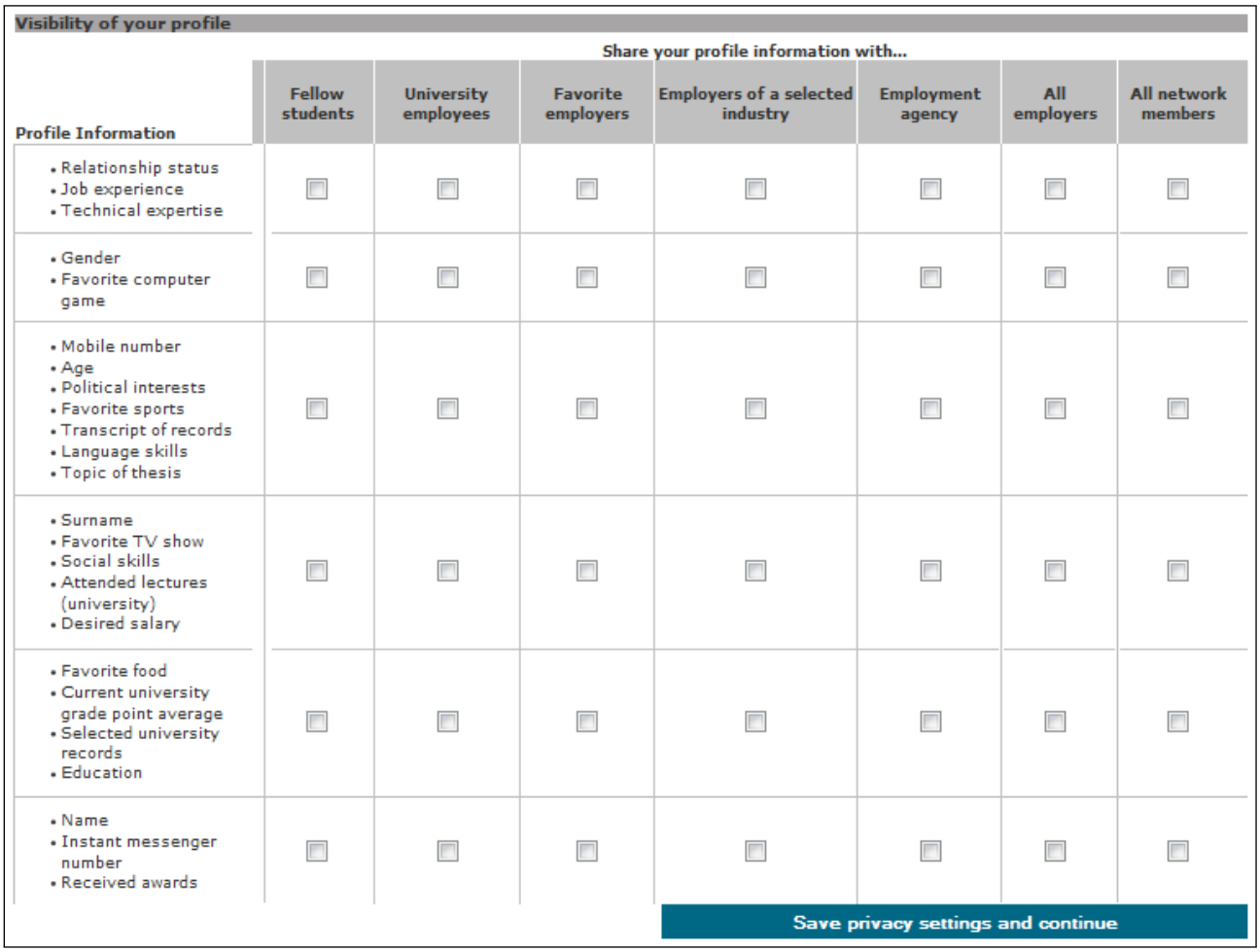


Figure 11: ACM layout for Condition 3: large choice amount and heterogenous object structure. (Study 2)

Out of the Loop: How Automated Software Updates Cause Unintended Security Consequences

Rick Wash, Emilee Rader, Kami Vaniea, Michelle Rizor
Department of Media and Information
Michigan State University
{wash,emilee,vaniea,rizormic}@msu.edu

ABSTRACT

When security updates are not installed, or installed slowly, end users are at an increased risk for harm. To improve security, software designers have endeavored to remove the user from the software update loop. However, user involvement in software updates remains necessary; not all updates are wanted, and required reboots can negatively impact users. We used a multi-method approach to collect interview, survey, and computer log data from 37 Windows 7 users. We compared what the users think is happening on their computers (interview and survey data), what users want to happen on their computer (interview and survey data), and what was actually going on (log data). We found that 28 out of our 37 participants had a misunderstanding about what was happening on their computer, and that over half of the participants could not execute their intentions for computer management.

1. INTRODUCTION

Home computer software is rarely released problem-free; most companies release a number of *software updates* to fix bugs in the software and add in new features. Microsoft alone released over 300 distinct software updates in the first three months of 2013. *Security* updates are particularly important because they are one of the primary mechanisms for protecting home computers from malicious software that leverages known vulnerabilities. The majority of computer compromises result from vulnerabilities for which a security update is available but has not yet been installed [16, 19]. Timely installation of security updates can protect users from the most common attacks [19].

Since installing security updates is so important for computer safety, many software companies have worked to find ways to improve end-user compliance and increase the number of fully updated systems. For example, each successive version of Microsoft Windows has had additional features to automate the installation of software updates with less human involvement [10]. Current software updates (and Microsoft Windows Updates in particular) have largely removed the need for human decisions. They default to automatically downloading and installing updates in the background, and forcing users to reboot (if needed).

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

However, not all security technologies can completely remove the human from the decision-making process [1]. Cranor assembled a useful framework for reasoning about when it is advisable to keep a 'human in the loop' [5]. This framework is relevant to software updates because updates cannot be installed completely without user intervention for three reasons: 1) occasionally, an update will introduce a new bug into the system, and users will want to postpone installing that update, 2) an update may introduce or remove features which impact user activities causing users to want to avoid installing the update, and 3) many updates require rebooting the computer to install, which is highly disruptive of user activities. Therefore, users need to be kept informed and given options during the update process. Software update systems have tried to accomodate users by finding an appropriate balance between forcing users to install updates to improve security, and giving them appropriate choices.

We conducted a multi-method user study to better understand how people make decisions about software updates that are so crucial to security. With each subject, we conducted semi-structured interviews to understand how the subject views software updates, had him or her take a survey to provide more structured opinions, and collected log data about update installation from his or her computer. In this paper, we focus primarily on subjects' decisions and behavior for Microsoft Windows updates. We find that over half of our subjects were not aware of what their computer's software update settings were or when the software updates were being installed. The majority of users' computers behaved in a way contrary to the user's intentions. However, many of these computers were also more secure than the user intended. This means that improving usability of software updates might not lead to improved security, which has interesting implications for the design of software update systems.

2. INTEGRATING HUMANS INTO SECURITY

Security failures are often seen as a human problem rather than a technological one. For example, West [24] wrote, "The most elegant and intuitively designed interface does not improve security if users ignore warnings, choose poor settings, or unintentionally subvert corporate policies."

In the workplace, computer and information security is the joint responsibility of end users and system administrators, but end users are often seen as "inherently insecure" [1, 11]. With the rise of discretionary computer usage and "bring your own device," end users bear more of the responsibility for the security of their many devices in and out of the workplace. Such users are their own system administrators, whether they know it or not, and how to best support them is the subject of much research.

Users are perceived as the weak link for several reasons:

- The expectations placed on end users with respect to managing the security of their computers are unrealistic; users cannot be expected to think like system administrators [2]
- Security only becomes apparent to end users when something has already gone wrong [27]
- Security is not users' first priority, and given a choice, they will choose the insecure option if it gets them closer to their goals [8]
- When users make mistakes, it makes the job of system administrators that much harder [8]

System designers frequently attempt to either nudge [20] or force users into making secure decisions. The designer might try to make security the user's top priority by creating mechanisms that prevent them from completing any action until the security aspects have been taken care of. The system might make the security-related actions so easy and unobtrusive that they can do whatever is necessary as part of their normal workflow or primary task (path of least resistance). Or, it might remove all responsibility and ability to act from the user by completely automating the security aspects of the system, so users cannot make the wrong choice [26].

However, it isn't feasible to completely automate security. Human capabilities are frequently necessary for the task at hand [22]. A "default" level of security is not appropriate for all users in all situations [9]. And automatic security cannot be used when configuration decisions must be made, or when automation is too "restrictive, inconvenient, expensive, or slow" [9]. Cranor [5] advocates that system designers should explicitly design for both automation and user responsibility for security by identifying which security aspects of the system cannot be automated and are likely to fail due to user intervention. System designers should provide better support to the users in those circumstances.

Software designers need to be aware that there is a tradeoff between visibility and intrusiveness. In circumstances when the user must remain "in the loop", communication between the system and the user is crucial, and it is the role of the software designer responsible for making sure the software is secure to figure out where this communication must take place [5]. Relegating security to "Advanced" tabs and burying it in menus is one way to (intentionally or unintentionally) ensure that the user retains the defaults. [9]

How that communication might best be accomplished is the subject of much usable security research. One of the core values of usability is "walk up and use" interfaces that do not require special learning or expertise; however, this approach may result in prioritizing the usability aspects of the system over the security aspects, because security may be more complicated than a "walk up and use" interface can communicate [12]. Recommendations to improve the usability of the communication between the system and the user are often assumed to also improve security, because users will be more involved, but this is not always the case.

To further complicate matters, end users often delegate the responsibility for the security of their systems, to technology, other people, organizations, or institutions [7]. Delegating responsibility to technology—to the system itself—is like 'set it and forget it' security: do it once, and never have think about it again. Once this has taken place, security becomes invisible, and is not often revisited. This means that the system keeps going with the past settings indefinitely. Policies like this are too rigid, because an invisible policy can't adapt to users' changing needs and circumstances [8].

Software updates are a particularly interesting case for studying how to include humans in security systems. From a security perspective, quickly installing security updates is the correct behavior, and can often be safely initiated without user intervention. However, many updates require that the computer reboot to complete installation, necessitating human involvement, and making the automated update process visible to users who may not understand why it is necessary [21].

3. SOFTWARE UPDATES IMPROVE SECURITY

Updating software is an important part of keeping a computer secure, and keeping all software up-to-date will protect a user against the most common security exploits. Symantec has data showing that the majority of computers are compromised using vulnerabilities where an update is available, but has not yet been applied [19]. The majority of web exploits use the top twenty vulnerabilities, all of which have available updates [19]. Likewise, Microsoft observes that all of the vulnerabilities exploited by the most popular exploit kit have available updates [16].

It is important to update software as soon as possible after a security update is released. Updates correcting security vulnerabilities are released an average of 1.2 months after an exploit for the vulnerability seen in the wild [15]. However, exploits released before a vulnerability becomes public knowledge (zero-day vulnerabilities) are used to attack a relatively small number of computer systems. Once a zero-day vulnerability becomes public knowledge the number of exploits using it increases 183–85,000 times and the number of attacks increases 2–100,000 times [3]. Likely for this reason, 60% of Microsoft's vulnerabilities are made public knowledge the same day as the update correcting the vulnerability is released [15], enabling users to protect themselves before exploits become readily available. For these and other security reasons, the faster the user updates their system the less likely they will be vulnerable to new attacks.

While updating quickly is good for security, all updates cannot be completely automated because they impact end users' workflows [21]. Many software updates include new, unwanted features. Some software updates introduce new bugs or incompatibilities. Rebooting interrupts users from their work. And many users prefer to "not fix what ain't broken."

There has been limited investigation into what motivates users to update or not update software on their computer. LaRose et al. surveyed undergraduate students about their online safety behaviors and beliefs. They found that people who feel like online safety is their personal responsibility are more likely to want to perform safe online behaviors [13, 14]. They also found that coping efficacy beliefs were correlated with intention to perform software updates [13]. These studies are based on self-report data, and are unable to examine whether subjects actually undertake their stated behaviors.

3.1 Windows Update

In this paper, we focus on Windows Update, a software update service provided for free by Microsoft. It began as a website that Windows 95 users had to visit to find out whether operating system updates were available. A new "Critical Update Installation Tool", introduced with Windows 98, included automatic checking for updates, and it also notified users about critical updates which they had to then manually retrieve and install. In 2000, Windows ME shipped with "Automatic Updates", a tool that could automatically download and optionally install software updates. Automatic installation of updates became the default with Windows XP SP2, and Windows Vista began automatically installing both updates categorized as "important" (including 'security' and 'critical' updates

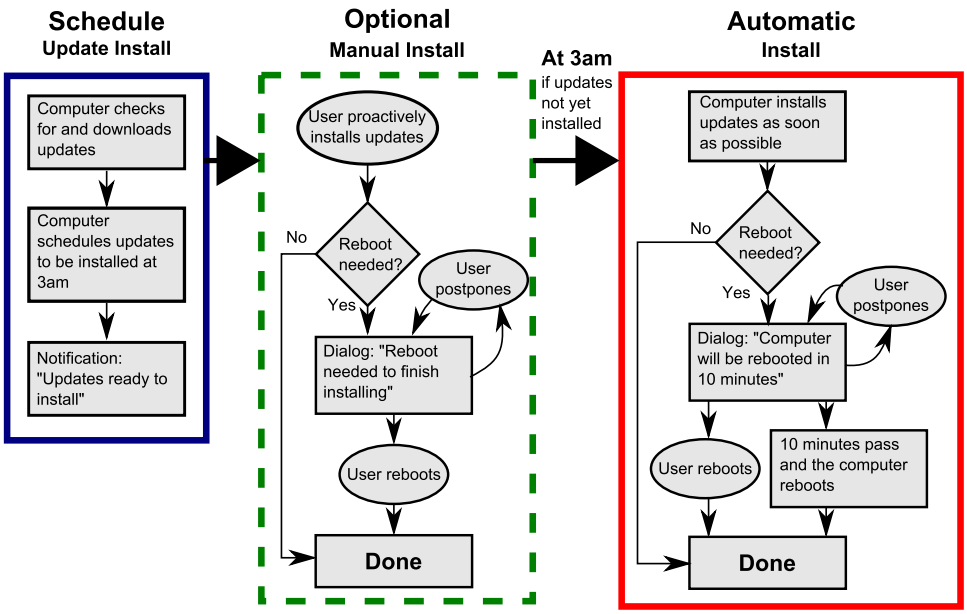


Figure 1: The Windows Update process. Ovals represent user actions, diamonds decisions, and rectangles computer behavior. This diagram was created based on prior update work by Gkantsidis et al., and experimentation using a Virtual Machine with Microsoft Windows 7 Service Pack 1 installed.

as well as reliability improvements), and also "recommended" updates [25].

The result of this evolution, the Windows Update software used in Windows 7, demonstrates the compromise Microsoft software designers made between automating the update process for the safety of users and giving users responsibility for their computer use. As shown in Figure 1, by default each update in Windows Update goes through three stages: an install scheduling, a time for manual install, and an automatic installation.

Stage 1: (left blue box) The computer automatically checks for updates, downloads them, schedules them to be installed at 3am the next morning, and then notifies the user that updates are available to be installed. The notification appears temporarily in the bottom right of the screen, and a gold shield is added to the "Shut down" button on the start menu.

Stage 2: (green middle box) The computer waits silently for the user to manually initiate the install process. This gives the user the opportunity to take responsibility for their updates. Users may manually install updates by opening the Windows Update program and selecting "Install updates." If a reboot is needed, the user is notified by a dialog with a postpone option. However, the dialog only reminds the user, it does not compel a reboot.

Stage 3: (red right box) The computer starts installing updates automatically at 3am or the first time the computer is turned on after 3am. If any update requires a reboot the computer presents the user with a dialog warning that the reboot will happen in 10 minutes. The dialog countdown timer has options to "Reboot now" or "Postpone"; the user cannot escape the countdown completely. If the user does nothing, the computer will immediately reboot. However, if the user chooses to intervene during the 10 minute interval, they can "Restart now" which causes an immediate reboot of the system, or "Postpone" for an additional 10 minutes, 1 hour, or 4 hours. This stage automates security decisions, removing the human from the loop.

The design of Windows Update is a compromise between fully automating updates and giving users full responsibility for updates, and it has been successful at increasing security. After the release of Windows XP SP2, Gkantsidis et al. observed that only 5% of SP1 users had fully updated computers, but 90% of SP2 users had fully updated computers. They also observed that 80% of SP2 users downloaded the latest update within two days of release [10]. In 2011, 66% of Windows users (all versions) were completely up-to-date, and 84% had at least one of the three most recent updates [16].

4. METHODS

Software updates are an instance where security system designers have mostly, but not completely, removed humans from security decision-making. To better understand user decision-making about software updates, we undertook a multi-method study that included semi-structured interviews, an online survey, and log-data analysis. This allowed us to measure both users' beliefs and impressions about what their computers were doing, and what their computers were actually doing.

4.1 Participants and Protocol

To study software updates, we wanted a population that doesn't have formal security or computer administration training, but still thinks enough about issues around updates that they have relatively well-formed opinions. We chose to study graduate students at a large research university in the Midwest of the United States. Graduate students are a group of computer users who are mostly non-technical, are responsible for maintaining their own computers, and depend on their computers for their work.

We sent an email through the University Registrar to a random sample of 1000 graduate students, excluding Math and Engineering students, asking for volunteers to participate in the study. Ninety-five people took a screening survey to ensure that they were Windows 7 users (so we could collect log data) and did not have any formal training in computer management, IT, or system administration. For this study, we chose to go deep into a single system's

updates, and chose the most popular, and most commonly exploited end-user system (Windows) to focus on. Thirty-seven people who were eligible came to our lab to participate, and brought their laptop running Windows 7 with them. Three of these subjects were Mac users running Windows in a virtual machine. Participants ranged in age from 21 to 57 with an average age of 31; Seventeen were male, and twenty were female. These demographics approximately match those of the larger graduate student population.

After informed consent, the study consisted of three parts: a brief survey, Windows log data collection, and a semi-structured interview. While one member of the research team administered the survey and interview, another member used a custom Powershell script to collect setting and log data on the subject's laptop. Subjects were given the option of observing the data collection. This study was approved by our university's IRB.

4.2 Three Types of Data

We collected three different datasets from each participant: a set of survey responses, log data from their Windows 7 laptop, and a transcribed, semi-structured interview. We began by analyzing each type of data separately. Then, using an ID number and pseudonym assigned to each subject, we re-combined the three data sources to compare subject responses and behavior across data sources. This analysis structure ensured that we accurately understood the meaning of each separate type of data before comparing attitude, recall, and behavior across data sources.

4.2.1 Semi-Structured Interviews

System designers have made most software updates highly automated and relatively invisible to end users. Users don't spend much time thinking about software updates. This poses a challenge for conducting interviews: how can we get subjects to talk about past experiences and reveal how they think about updates? And how can we avoid having subjects think about updates too deeply during the interview – and change their opinions, which would lead to invalid data?

After a series of pilot tests, we decided to use three interviewing techniques: free-listing, hypothetical scenarios, and recollection of specific past instances.

We began by asking participants to complete a *free-listing activity* [4]: write down as many examples as came to mind for the prompt, "things that can happen if the software on your computer is too old or out of date". We then read each example and asked the participant to discuss his or her response further. Free-listing allows us to explore the semantic domain of updates; that is, it helps the subject to think through and explain the range of activities and concerns that are relevant to a discussion of software updates. The use of a non-specific prompt, reading items back to the subject, and using the items as semantic cues to discuss past instances help subjects to fully explore the topic of software updates [4].

Next, we presented subjects with a series of five *hypothetical scenarios* paired with probing questions; we wanted the participant to do most of the talking so that we could uncover their attitudes, beliefs, and mental models about updates. The scenarios involved being prompted to restart an internet browser mid-task, seeing that a large number of urgent Windows updates were available, reading a news article about a virus, a software program that costs money to update, and a slow computer with lots of warnings. Hypothetical scenarios are effective methods of learning how subjects conceptualize their decisions relate to software updates [23].

Finally, throughout the interview, we regularly asked subjects to *recall specific past instances* of software update decisions. By asking to recall specific instances, subjects provide more details and are better able to recall information that influenced their decision-making at the time. Recalling specific instances provides data that is more likely to represent broad decision-making patterns than asking subjects to describe general patterns of past behavior [18].

Analysis: After transcribing and anonymizing the interviews, we performed a bottom-up, inductive coding. We started with an initial list of themes identified by the research team, and expanded the codes as each of us separately read through transcripts. During this period, members of the team met frequently to discuss and revise the codes. Themes identified include "negative update experiences", "attitudes toward delaying updates", and "why updates are important."

As we created each code, we examined other subjects to check for representativeness and identify which traits were common across subjects. We also explicitly looked for negative cases: cases that share most of the pattern but are explicitly missing one or two key pieces.

When coding was complete, we summarized the data into a matrix that displayed themes by participant [17]. This matrix allowed us to understand each individual's perspective on updates by reading down the column that summarizes their responses. We then compared the summary data matrix to original interviews to verify the correctness of each summary, check for the meaning of outliers, verify surprises, specifically look for evidence for negative cases, and try to prevent researcher confirmation bias in our data. [18]. This process provides confidence that our summaries are valid representations of participant views as expressed in the interviews.

4.2.2 Survey

We used an in-lab computer survey to ask structured, closed-ended questions. A survey allowed us to ensure that all participants were asked the same set of factual and opinion-based questions in a consistent, comparable manner. In addition to background information such as subject demographics, computer skills, and installed software, we also asked subjects for their current understanding of the state of software updates on their computer. This includes whether automatic updates were enabled and whether updates were usually installed manually or automatically. Questions were written following the guidance of Dillman [6] and were pretested to ensure subjects understood the questions the same way the researchers did.

Analysis: We generated descriptive statistics for each subject, as well as extracting the specific questions about the user's knowledge of current state of the automatic updates setting, their belief about whether updates are installed manually or automatically, and their belief about the timing of install. The full survey instrument is available in the Appendix.

4.2.3 Windows Logs

The Windows operating system, along with many Windows services, records information about system events in log files which contain detailed records of system and user behavior. Our Powershell script collected the current Windows Update settings, which allowed us to determine whether updates were turned off, set to notify the user before download, or set to install automatically without user intervention (default behavior). The script did not collect any personally identifiable information.

We also collected a list of installed updates from the Windows Update API, and a copy of all Windows Update log files which provided detailed event information from the last several months of use. This allowed us to calculate the time between when an update had been downloaded and when it was installed, which is important because this is the part of the update process that the user has

the most control over—i.e., when the update is installed and when the computer reboots to finish installing an update (if necessary). One limitation of this method is that the detailed logs represented between 1 and 17 months (average of 6) of usage data depending on how often the participant had been using the machine.

Analysis: We first looked at each update separately. We limited our log analysis to updates which were associated with a Microsoft Knowledge Base (KB) number, which allowed us to link update events across log files. We marked the update as proactively installed by the user if it was installed before 3am[1] the morning following the update's download. We marked it as automatically installed by Windows Update if it was installed after 3am. Then we aggregated all updates for a user: did the user always install proactively (100%), usually (> 50%) install proactively, usually automatic install, or always automatic install?

4.3 Combining Data for Analysis

In order to compare user attitudes, user beliefs, and user behavior, we constructed a data matrix that combined data from all three sources of information [17]. For each subject, we created entries on three topics: general updates, the automatic updates setting, and the timing of update installs. For each of these topics, we included a row of data from each of the three data sources: the subject's attitude and understanding of the topic summarized from the interviews, the subjects current beliefs from the survey, and the subject's past behavior summarized from the log data.

After creating the combined data matrix, we again examined our data to ensure validity [18]. All members of the research team participated in looking for patterns across subjects, checking for negative cases, verifying summaries with original source data, and including footnotes and caveats for our summaries.

For each of the three topics, this data matrix allowed us to directly compare a subject's understanding, the subject's belief, and the subject's behavior on their computer. In checking through this data matrix, however, we noticed that subjects' understanding and beliefs were not straightforward. Rather, each subject's understanding and beliefs could be separated into two: the subject's understanding of what his or her computer is currently doing, and the subject's intention for what he or she would like the computer to be doing. Therefore, we split these understanding rows in two, and verified each piece with the source data.

5. FINDINGS

We used our interview data and our survey data to characterize two things: what the user thought the computer was doing, and what the user wanted the computer to do. We then compared these two perceptions with the log data from that user's computer to determine if they matched. That is, we compared user's stated *understanding* of what their computer was doing with log data and settings that indicated what the computer actually did, to see whether users understood what was happening on their computer. Then we compared each user's stated *intentions* — what they wanted their computer to be doing — to the log data and settings to determine whether they were actually able to make the computer do what they wanted.

5.1 Understanding Software Updates

Many of our subjects misunderstood what their computers were doing regarding software updates. Twenty-eight of the 37 subjects (78%) had at least one inconsistency between what the subject thought their computer was doing and what the log data indicated it was doing. There are two topics that subjects had misunderstandings about: the Windows Update setting about whether to install updates automatically, and how quickly updates were installed.

Consistent		*Inconsistent*	
Changed Setting	4	On, but thinks Off	4
Default Setting	8	Off, but thinks On	2
		Download but not Install	5
		Notify, but not Download	14
Total	12	Total	25

Table 1: Misunderstandings of Automatic Updates (Number of Subjects)

Automatic Updates Setting.

Automatic update settings were a prevalent source of misunderstanding for our subjects. There are four possible settings in Windows Update: 1) *On*, the default setting where Windows automatically downloads and installs updates according to the process described in Section 3.1 (31 participants had this setting), 2) *Download* available updates but do not install them (0 participants), 3) *Notify* the user when updates are available, but do not automatically download or install them (4 participants), and 4) *Off*, where Windows Update must be manually run for anything to happen (2 participants).

Among our 37 subjects, 25 had some form of inconsistency between what they stated they thought their computer's auto-update setting was, and the recorded settings on the computer (See Table 1). Of these, five subjects were close to correct: they thought that their computer automatically *downloaded* updates and prompted them to install. While this is true, their actual setting automatically installs the downloaded updates at 3am if the user hasn't already installed them; these five subjects frequently installed their updates proactively so rarely encountered the 3am automatic install.

This leaves 20 subjects who had an inconsistency in their understanding of their auto-update setting. Four subjects believed that their auto-updates had been turned off, when in reality they had the default, secure setting of automatically installing updates. Two subjects believed the opposite; they thought they had auto-updates turned on, but auto-updates had been disabled on their computer[2]. The remaining 14 subjects expressed a belief that automatic updates only notify them about available updates but do not install them. However, these 14 subjects all had the default setting of automatically installing updates. For example, Justin[3] told us "I mean it usually prompts me when there is an update to be installed, but I don't know if that means auto-update or not." His survey answers also indicated that he thought that Windows notified him, but did not install updates.

As a comparison case, 12 subjects were completely consistent in their understanding of auto-updates. Eight had the default setting, and correctly understood that setting as automatically downloading and installing updates. Rachel said, " I guess my current belief is that the operating system doesn't give you a choice about updating things, it just does it for you." And four subjects had intentionally changed the setting to *Notify Before Download* (i.e., the computer notifies the user that new updates are available but does not down-

[1] One user had a scheduled install time setting of 4am, all other users had the default of 3am, for simplicity we always refer to this time using the default of 3am or "overnight".

[2] One of these subjects may be running a third-party updating system designed for pirated Windows systems.

[3] All subject names have been anonymized.

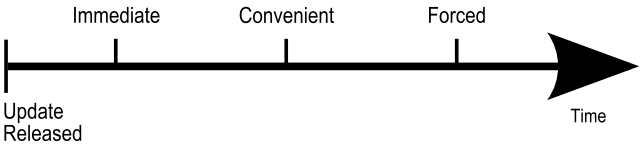


Figure 2: Perceived Times When Updates Can Be Installed

load or install them), and also correctly understood their change.

In our sample of non-technical computer users, six subjects' computers did not have the default auto-updates setting, *Scheduled Install*, in which software updates are automated as much as possible. Two of these subjects didn't understand the setting and thought they were still on. However, the remaining four subjects correctly understood that their computers would not automatically install updates. An additional 14 subjects, who had the default setting of *Scheduled Install*, believed that they were only notified about updates and that no updates were installed automatically. These findings indicate that many misunderstandings exist regarding whether users are updating Windows, and that sometimes these misunderstandings mean that updates are not installed.

Timing of Update Installation.

The timing of updates is another source of inconsistency between subjects' stated intention and log data. Common security advice is that software updates, and particularly security updates, should be installed as quickly as possible to protect against in-the-wild exploits and zero-day vulnerabilities [19]. However, installing software updates usually interrupts what the user is doing on their computer, and often requires a severely disruptive reboot [21].

In our log data analysis, we characterized each update as either *proactive* or *automatic* depending on if the user proactively installed the update, or if Windows automatically installed the update. Each subject, then, made a series of choices that either resulted in the subject installing most of their updates proactively, or mostly allowing Windows to automatically install.

However, subject understanding of update timing doesn't exactly match this characterization. Instead, we found three timing categories for when updates might be installed (See Figure 2). The fastest possible update installation happens when a user is notified about an available update, and interrupts what they are doing to *immediately* and manually install the update. An intermediate timing occurs when a user is notified about an update, but doesn't interrupt their work to install it immediately. Instead, they wait until a *convenient* time to manually install the update. Both these categories involve manual installation, though some users may not find convenient times and end up with Windows automatically installing some updates. Finally, the slowest timing that actually results in the update being installed corresponds with the *forced* timing, and occurs when the user waits too long and the computer automatically installs the update and reboots the computer.

This difference in technical coding and user understanding poses an analysis challenge: when a subject indicates that they install their updates "when convenient," how do we characterize whether their behavior is consistent with their understanding? To address this, we first looked at the logs for whether most of an individual subject's updates were automatic or manually installed. If updates were mostly automatic, then that is a clear disconnect from the subject's stated understanding of installing when convenient; since the automatic install happens as pre-specified times, it is unlikely that that is happening "when convenient."

However, if the subject mostly installed updates manually, then this could be consistent with a desire for convenience (if they waited until it was convenient to install and reboot), or it could be inconsistent (if they interrupted themselves to install the updates). Since whether or not a subject was interrupted is entirely in the opinion of that subject, we looked to the survey data for guidance on how to categorize them. On the survey, we asked each subject how likely they would be to interrupt themselves to install Windows updates. Consistent with traditional interpretations of similar Likert scale survey questions [6], we took this question to represent the subject's memories of whether they were frequently interrupting their work to install updates. If they answered "Likely" or "Very Likely", then we took this as inconsistent with their stated desire for convenience. Any other answer was considered consistent.

Consistent		*Inconsistent*	
When Convenient	8	Want Convenience, but Automatic	8
		Want Convenience, but Proactive	6
Wait till Forced	6	Thinks Delay, Installs Proactively	2
		Wants Only AV updates	2
		Turned Auto-updates Off	1
Total	14	Total	19

Table 2: Inconsistencies in Timing of Update (Number of Subjects). We excluded four subjects from the table due to insufficient information.

Results: Nineteen of our 37 subjects expressed a desire about the timing of updates that was inconsistent with the log data on their computer. Of these, ten subjects installed updates more quickly than their stated intention, and nine subjects installed updates more slowly. (See Table 2 for counts.) Four subjects had insufficient interview data to accurately judge their desires.

Twenty-two subjects stated that they wanted to install updates manually at a convenient time; however, eight of them never actually got around to running the updates and the computer ended up automatically installing the update — which means the subjects installed updates slower than intended. Six subjects actually interrupted their work and installed the updates very quickly. On the diagram in Figure 2, all 22 of these subjects' stated intentions were to install in that middle range of timing — when convenient. Eight actually installed at that time; eight actually installed when forced (to the right), and six actually installed immediately (to the left).

Two subjects stated that they usually delay updates, particularly updates that require a restart. These subjects, however, usually installed updates very quickly according to the logs. Three subjects said they only do updates labeled "urgent"; two of them successfully installed all updates quickly, but one subject had auto-updates turned off and didn't install any updates.

When a subject has an inconsistency about when updates are being installed, this isn't a technical misunderstanding. Subjects aren't misunderstanding how the computer is working. Rather, they are misunderstanding their own behavior. Such a misunderstanding is important because it can form the basis for further decisions, such as "is my computer secure?" But since it is not a technical misunderstanding, greater education will not necessarily solve it.

Difficulty Understanding Updates.

As indicated by the many inconsistencies mentioned above, many of our subjects misunderstood what was happening on their computers. In examining our interview data, we found two reasons they were having problems.

First, the computer wasn't very clear about what it is doing and

when it is doing it. Many subjects talked about how it was difficult to understand what was going on. Nicole, for example, could not tell whether she permitted her computer to automatically update or not:

> Actually I didn't know that I clicked yes for auto updating. It just popped up. So, that's why I know about the auto updating. And other stuff, I didn't know that I clicked yes for auto updating or something like that.

In the interview, she indicated that she thought it was important to install urgent and critical updates, and in the survey she indicated that she thought her updates were automatically installed. However, her computer actually had automatic updates turned off.

Second, even when our subjects tried to look at settings and dig deeper, they found most of the settings to be confusing and difficult to use. Matt said that he "[doesn't] even know where I'd go to do that." Will wanted to turn off automatic updates:

> But I know I played around with some of the settings on my computer so that it wouldn't automatically update everything. Because it would just slow down my computer to a crawl. And several computers that I've had, it makes it harder when you're trying to get a task done.

However, Will's computer still had the default setting and all updates released had been installed. Furthermore, most of his updates were automatic installs, rather than being installed manually.

Many of these misunderstandings stem from design choices that try to remove the need for humans to make decisions about software updates. Windows Update has automated as much as possible and moved many updates actions into background, invisible processes. That automation made it difficult for many of our subjects to understand what was happening on their computer at any time, and even whether updates were being installed at all. Additionally, to discourage users from changing settings, Windows Update makes it difficult for users to find the settings in the first place. So even if our subjects did want to change the settings, they couldn't figure out how. Removing the subjects' decision-making ability had the side effect of also making it difficult for them to learn about updates and understand what their computers were doing.

5.2 Intentions and Security

In addition to describing their current understanding, our subjects also described what they wanted to be doing about software updates. Did our subjects intend to put off updates because they felt like updates weren't important, or did they intend to install them immediately but ended up delaying indefinitely? Here, we describe whether these stated *intentions* match what was actually happening on the computer. Mismatches between intentions and behavior indicate usability problems, or what would change if we made software updates easier to understand and use.

For this analysis, we consider installing updates to be secure, and installing them sooner is more secure than waiting and installing them later. While users may have good reasons to choose to be less secure, we focus primarily on the security consequences of those choices.

Two subjects provided short answers during their interviews and did not clearly describe their intentions for what they wanted their computers to be doing. Therefore, these subjects were removed from this analysis of intentions.

Consistent		*Inconsistent*	
Notify but not Auto-Install	3	More Secure	12
Not urgent, so wait till Forced	3	Less Secure	9
Always install Immediately	8		
Total	14	Total	21

Table 3: Whether Intentions are Consistent with Reality (Number of Subjects)

When Intentions Don't Match Reality.

Twenty one subjects had a disconnect between their stated intentions for installing software updates and what the log data indicated their computer was actually doing (Table 3).

For nine of these subjects, the computer ended up being less secure than the subject intended. Three subjects intended to install updates regularly and automatically, but actually had their automatic updates turned off (or to notify) and had almost no updates installed on their computer. The remaining six subjects all stated that they intended to proactively install updates as soon as it was convenient, but rarely actually got around to installing the updates until the computer automatically did so. This mismatch between intention and behavior led to the updates being installed, but left a larger window of vulnerability than the subject intended.

As an example, Dan talked about how he chose when to install updates:

> If I were doing something fun I would interrupt it, no problem. If I were just surfing the web, it's like, oh, whatever, I'll update my computer. But if I'm writing an email, if I'm working on a paper, if I'm working on a homework assignment, then that usually takes priority. If I can put it off for 15, 20 minutes, I'll just do that later then, 'cause when I'm in the zone studying, I don't wanna be interrupted with anything.

This is a typical representation of a "convenient" intention: he wanted to install updates, but didn't want to be interrupted. So he said he'd finish what he was doing and then install the updates. However, Dan's computer logs indicated that Windows Update automatically installed most updates; he rarely installed them manually. This means that his computer was vulnerable for the maximum amount of time that Windows Update allows.

Twelve subjects had a disconnect between their stated intentions and the log data that left their computer more secure than they had intended. Two of these users explicitly stated that they wanted to turn automatic updates off, but their computer still had the default setting of automatically downloading and installing updates. Another example is a subject who wanted to continuously delay updates, indefinitely, but had the default auto-update setting that automatically installed updates in a relatively timely fashion.

One subject from this group, James, expressed an intention to delay updates until a convenient time, but always ended up interrupting what he was doing to manually install updates. He described one instance that illustrated his intention to install when "convenient":

> What was I gonna do? I was working on homework for something and I was loading a video on my browser to watch while I ate food. It was buffering and loading, and I usually will take a meal break and watch a movie at the same time. And I realized if I restarted, then that would have to reload, the movie would have to reload

> all the way from the beginning. And I would lose that time because I was going to eat in 15 or 20 minutes and then I had to go somewhere, I had a class. So I decided, you know what, I'll just postpone.

However, according to James's computer logs, all of the updates on his computer were installed, and were installed manually in less than 24 hours after being downloaded. James actually interrupted his computer use at some point rather than postponing, and ended up with a smaller window of vulnerability than he would have if he had waited to install when convenient.

These disconnects are interesting when we look at what would happen if we improved the usability of software updates and did a better job of including the user in the loop. Nine of our subjects' computers would be more secure if they were able to execute on their intentions, while twelve would be less secure. The sample for this study is not representative, so we cannot claim that these 21 out of 37 subjects (59%) generalize to the larger population of computer users. However, our sample has a relatively large number of both people who would be more secure if usability improved, and a similar number who would be less secure if usability improved. We suspect that both groups are well-represented in the larger population.

When Intentions Match Reality.

Fourteen of our subjects were able to successfully execute on their intentions: the log data from their computer was consistent with these subjects' stated intentions for software updates. However, these subjects had varying levels of security.

Eight subjects fell into the most secure category; these subjects all had the default setting that automatically downloads and installs updates. These subjects felt strongly that installing updates is important, and manually installed updates soon after they were notified that the updates were available. These subjects didn't wait for the computer to automatically install the update. By manually installing the update, they minimized the window of vulnerability.

Three subjects had a strong objection to the way that Windows compels the computer to reboot; these subjects felt rebooting seriously interrupted their work. These subjects changed their settings so that Windows notified them that updates were available, but did not download or install them. They manually downloaded and installed updates at a convenient time. Everyone in our study who had changed their auto-update setting to *Notify Before Download* or *Notify Before Install* fell into this group; people who change this setting seemed to understand that updates are important and still install them, but not as quickly.

Finally, three subjects didn't feel like updates were that important, and wanted to have the computer deal with the updates for them. They continually postponed updates until the computer automatically installed the updates, and rebooted their computer.

Would Better Usability Be More Secure?.

Many people in the HCI community emphasize usability; if we make computers easy to walk up and use, then people will be able to accomplish more with them. When people form intentions about what they want their computer to do, but cannot execute on those intentions, HCI professionals naturally suspect a usability problem. Indeed, Windows Update seems to have a usability issue; 21 of our 37 subjects (approximately 59%) were not able to use the system the way they wanted to.

However, it isn't clear whether better usability would actually be an improvement in this case. Only 9 of 21 subjects whose behavior did not match their intentions were less secure than they wanted to be; these subjects would end up more secure if we were to improve usability. But for the remaining 13 subjects whose behavior did not match their intentions, the computer was more secure than it would be if usability were improved. These subjects wanted to be less secure, and poor usability was preventing them from executing on that intention.

Many of our subjects had misunderstandings about what their computer was doing with software updates. And many of our subjects had trouble executing on their intentions. One reasonable assumption is that the second statement — the difficulty in executing on intentions — is caused by the first. However, we don't believe this is the case. A couple of subjects completely understood what their computer was doing, but still could not execute on their intentions. For example, Rachel understood that the computer was installing updates, but felt like auto-updates were controlling her and forcing her to install them. And there were many subjects who didn't understand what their computer was doing, but ended up doing exactly what they wanted to. Brittany believed that her computer only notified her but didn't install updates; however, she wanted to control her updates and ended up installing almost all of her updates manually at convenient times. It seems that understanding is not necessary to be able to execute on security intentions.

6. DISCUSSION

Our subjects had a number of misunderstandings about what their computers were doing with respect to software updates. Also, our subjects frequently were not able to execute on their intentions about whether and when to install software updates. We speculate that these challenges may be the result of trying to remove the human from security decisions. We also observe that improving usability may actually backfire.

Learning Through Decisions.

In designing security technologies, there is a tension between removing human decisions to automate security, and allowing the user the flexibility to make important choices [5]. The current version of Windows Update represents a compromise; most of the decisions about updates are made by the computer, removing the human from decision making. Many updates are downloaded and installed automatically, and Windows eventually automatically installs all downloaded updates even when they require a reboot. Some human decisions remain, particularly when they impact use of the computer, such as rebooting.

Removing the human from decisions, however, seems to have had an unintended side effect: users now find it difficult to understand what the computer is doing, and to correctly implement their part of the updates process. Having to make decisions as part of a security mechanism helps the user to learn how that mechanism works, what decisions are appropriate, and how to correctly execute those decisions. This learning may be direct, coming from feedback within the system. Or, this learning may be indirect learning, with the user seeking out the knowledge necessary to make better decisions.

Windows Updates has successfully automated so many security decisions that many users don't learn how to make intelligent security decisions about software updates. Instead, they struggle at understanding what their computer is doing, and often fail to execute even when they do make a decision.

This is important when some, but not all, security-relevant decisions can be automated. Removing the user from most of the decisions makes it more difficult for the user to intelligently make the remaining decisions that cannot be fully automated.

Designing Update Systems.

There is a fundamental tension here between learning and understanding what the computer is doing, and improving security by forcing the user to behave securely. It isn't clear which is a better strategy. Consider just the results in this paper: if usability were improved and users were able to accurately execute on their intentions, some users would end up less secure but many would end up more secure. The net effect on security isn't clear; it is possible that ignorance and inefficacy might be better for security than learning and usability.

There is also a tension here among the users. Some users want to trust the computer to make good decisions for them; that is, they want the computer to be its own system administrator. For these users, automating good decisions is valuable. However, other users want control over their computer, and rebel against the feeling of being forced into doing things they don't agree with (or just haven't thought about).

The software industry is currently struggling with these tensions. Windows update is clearly moving toward automating as much of the software update process as possible. A wide variety of other system applications are following. Firefox automatically downloads and installs updates with virtually no user intervention. Java is moving toward automatically installing updates, and Adobe is moving to a subscription model with automatically installed updates and upgrades. Apple's iOS 7 and OSX Mavericks now allow users to turn on a setting to automatically install updates to all software installed via the official App Stores.

However, some end-user "apps" and most business applications are moving to a much more explicit, user-driven update model. Some smartphones, for example, require the user to explicitly check for updates and choose to install them. Timing of this install is important. If you must pick a single install time, Windows did well. However, for any individual in a specific week, that time might not always be convenient. *Idle* on a computer does not necessarily mean *convenient* – it could be that users have important state that would be lost if an update was installed or the computer rebooted. A better strategy might be an adaptive mechanism that detects and when the user is finishing their work for the night and provides a notice at that time.

Almost all software on PCs eventually requires software updates, and many of these updates are security relevant. Each software vendor makes choices about how to distribute these updates. Our results suggest that automating updates similar to Windows Update or Firefox will lead to more uniform update installations, but will also result in many users not understanding what is happening on their computers and not being able to change things when they want to. On the other hand, manually installing updates may lead to better understanding about updates and greater feeling of control, but will also likely result in lower levels of security and compliance.

7. CONCLUSION

Quickly installing software updates is one of the best ways to protect your computer from malicious attackers. To improve security, companies such as Microsoft have moved to a model of automatic software updates that removes much of the decision-making by the end user. Using a combination of interviews, a survey, and log data, we compared what non-technical users understand about what their computer is doing to install software updates, what they want their computer to be doing, and what is actually happening on the computer.

We found that many end users had misunderstandings about what was happening on their computer; more than half our our subjects didn't correctly understand the automatic update settings on the computer, and more than half of our subjects did not understand when their updates were being installed. Furthermore, when users decided how they wanted to manage software updates, they often could not execute on that intention. This mismatch between intention and behavior frequently led to the computer being more secure, but also frequently led to the computer being less secure than intended.

8. ACKNOWLEDGMENTS

We thank Zack Girouard for his assistance with data collection and early analysis. We thank everyone associated with the BITLab at MSU for helpful discussions and feedback. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1116544 and CNS-1115926.

9. REFERENCES

[1] ADAMS, A., AND SASSE, M. A. Users are not the enemy. *Communications of the ACM 42*, 12 (1999), 41–46.

[2] BESNARD, D., AND ARIEF, B. Computer security impaired by legitimate users. *Computers & Security 23*, 3 (2004), 253–264.

[3] BILGE, L., AND DUMITRAS, T. Before we knew it: An empirical study of zero-day attacks in the real world. In *Proceedings of the ACM Conference on Computer and Communications Security* (New York, NY, USA, 2012), pp. 833–844.

[4] BREWER, D. D. Supplementary interviewing techniques to maximize output in free listing tasks. *Field Methods 14*, 1 (2002), 108–118.

[5] CRANOR, L. F. A framework for reasoning about the human in the loop. In *Usability, Psychology, and Security (UPSEC)* (2008).

[6] DILLMAN, D. A., SMYTH, J. D., AND CHRISTIAN, L. M. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 3rd ed. Wiley, Hoboken, NJ, 2009.

[7] DOURISH, P., GRINTER, R. E., DELGADO DE LA FLOR, J., AND JOSEPH, M. Security in the wild: User strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing 8*, 6 (2004), 391–401.

[8] EDWARDS, W. K., POOLE, E. S., AND STOLL, J. Security automation considered harmful? In *Proceedings of the New Security Paradigms Workshop, NSPW* (2007), pp. 33–42.

[9] FURNELL, S. Why users cannot use security. *Computers & Security 24*, 4 (June 2005), 274–279.

[10] GKANTSIDIS, C., KARAGIANNIS, T., AND VOJNOVIC, M. Planet scale software updates. In *ACM SIGCOMM Computer Communication Review* (New York, New York, USA, Aug. 2006), ACM, pp. 423–434.

[11] KAEMER, S., AND CARAYON, P. Human errors and violations in computer and information security: The viewpoint of network administrators and security specialists. In *Applied Ergonomics* (2007), vol. 38, pp. 143–154.

[12] KAINDA, R., FLÉCHAIS, I., AND ROSCOE, A. W. Security and usability: Analysis and evaluation. In *International Conference on Availability, Reliability, and Security, ARES* (2010), IEEE, pp. 275–282.

[13] LAROSE, R., RIFON, N., LIU, S., AND LEE, D. Understanding online safety behavior: A multivariate model. In *The 55th Annual Conference of the International Communication Association* (New York City, 2005).

[14] LaRose, R., Rifon, N. J., and Enbody, R. Promoting personal responsibility for internet safety. *Communications of the ACM 51*, 3 (Mar. 2008), 71–76.

[15] Marconato, G., Nicomette, V., and Kaaniche, M. Security-related vulnerability life cycle analysis. In *Risk and Security of Internet and Systems (CRiSIS), 2012 7th International Conference on* (2012), pp. 1–8.

[16] Microsoft. Microsoft Security Intelligence Report, Volume 13, January – June 2012.

[17] Miles, M. B., Huberman, A. M., and Saldaña, J. *Qualitative Data Analysis*. A Methods Sourcebook. SAGE Publications, Incorporated, Apr. 2013.

[18] Onwuegbuzie, A. J., and Leech, N. L. Validity and qualitative research: an oxymoron? *Quality & Quantity 41*, 2 (2007), 233–249.

[19] Symantec Corporation. Internet Security Threat Report, Volume 18, 2013.

[20] Thaler, R., and Sunstein, C. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, 2008.

[21] Vaniea, K., Rader, E., and Wash, R. Betrayed by updates: How negative experiences affect future security. In *Proceedings of the ACM Conference on Human Factors in Computing (CHI)* (Toronto, Canada, 2014).

[22] von Ahn, L., Blum, M., Hopper, N. J., and Langford, J. CAPTCHA: Using hard ai problems for security. In *EUROCRYPT '03* (2003), pp. 294–311.

[23] Wash, R. Folk models of home computer security. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)* (2010).

[24] West, R. The Psychology of Security. *Communications of the ACM 51*, 4 (2008), 34–41.

[25] Wikipedia. Windows Update. `http://en.wikipedia.org/wiki/Windows_Update`; last retrieved September 17, 2013.

[26] Yee, K.-P. User interaction design for secure systems. In *International Conference on Information and Communications Security, ICICS* (2002), pp. 278–290.

[27] Zurko, M. E. User-Centered Security: Stepping Up to the Grand Challenge. In *21st Annual Computer Security Applications Conference (ACSAC'05)* (2005), IEEE, pp. 187–202.

APPENDIX

A. SURVEY QUESTIONS

Q1: Suppose there is a lottery where you have a 10% chance of winning $1000. What is the largest amount you would be willing to pay for a ticket in this lottery?

Q2: How do you see yourself: Are you in general a person who takes risk or do you try to evade risks? Please self-grade your choice (ranging between 0-10)

( ) 0 – not at all prepared to take risk
( ) 1
( ) 2
( ) 3
( ) 4
( ) 5
( ) 6
( ) 7
( ) 8
( ) 9
( ) 10 – very much prepared to take risks

Q3: How familiar are you with the following terms? Please rate your familiarity with each term below from None (no understanding) to Full (full understanding):

	None	Little	Some	Good	Full
Security Update	◯	◯	◯	◯	◯
Critical Update	◯	◯	◯	◯	◯
Service Pack	◯	◯	◯	◯	◯
Software Update	◯	◯	◯	◯	◯
Optional Update	◯	◯	◯	◯	◯
Hotfix	◯	◯	◯	◯	◯
Upgrade	◯	◯	◯	◯	◯

Q4a: Are you responsible for maintaining the laptop you brought with you today? Maintenance activities include things like installing and updating software, running antivirus, dealing with problems that may arise, etc.

( ) Yes
( ) No
( ) Other _____

Q4b: Is there another person (or people) who helps with maintaining the laptop you brought with you today?
(Shown only if participant is responsible for maintaining their laptop.)

( ) No, I do it by myself
( ) Yes, I share the responsibility with someone else
( ) Yes, I ask for help occasionally from someone who knows more than I do
( ) Other (please specify) _____

Q5: Please list the other people who use this computer, by their first name only. If nobody else uses this computer, leave the box blank:

Q6: Which of the following types of software do you have installed on the laptop you brought with you? Please check all that apply:

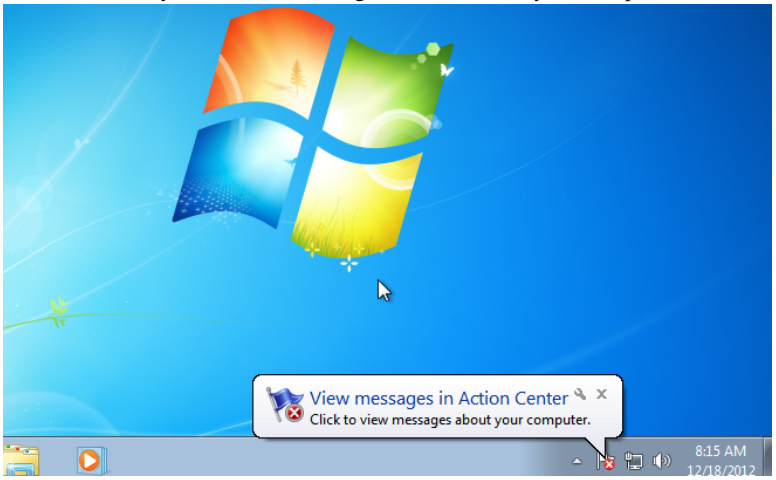
[ ] Windows operating system
[ ] Microsoft Office
[ ] Anti-virus software
[ ] Virus definitions or data files for your anti-virus software
[ ] Firewall software
[ ] Web browser, like Chrome or Firefox
[ ] Internet security software

[ ] Anti-spyware software
[ ] Adobe products, like Adobe Reader or Flash
[ ] Java
[ ] Database, like Oracle or Microsoft Access
[ ] Graphic design, like Photoshop
[ ] Multimedia, like iTunes, DVD player
[ ] Games
[ ] Communication, like Skype, Instant Message
[ ] Educational software

Q7b: Which of the following anti-virus programs do you have installed on your computer? Please check all that apply:
Only shown if the participant claimed to have an anti-virus installed.

[ ] Avast
[ ] AVG
[ ] Norton
[ ] McAfee
[ ] Microsoft
[ ] Kaspersky
[ ] I have an anti-virus program installed, but I don't remember which one
[ ] Other (please specify) _____

Q8: How often do you remember seeing a notification on your computer that looks similar to the following image?



( ) Never
( ) Rarely
( ) Sometimes
( ) Often
( ) Very Often

Q9: How long has it been since the last time any software on the laptop you brought with you was updated?

( ) Less than one month
( ) A couple of months
( ) 6 months or so
( ) About a year
( ) 1-2 years
( ) Longer than 2 years
( ) I don't know

Q10: In what ways do you remember finding out that a software update is available? Please check all that apply:

[ ] Checking the website of the software company

[ ] Checking for updates using the software itself

[ ] Email notification

[ ] News article

[ ] Mentioned by a friend or family member

[ ] Mentioned by a work colleague

[ ] Automated message on your computer

[ ] Other (please specify) _____

Q11:
Some kinds of software can check for software updates and let the user know when an update is available. Other kinds will check and then also download the update, so it is ready for the user to install. Still others automatically install software updates without any action by the user.
For each of the following kinds of software you indicated above that you have installed on the laptop you brought with you today, please indicate which kinds of software you remember behaving in the following ways:

CHECKING for updates automatically, and NOTIFYING you that new updates are available

CHECKING for and DOWNLOADING updates automatically, and NOTIFYING you that an update is ready to be installed

INSTALLING updates automatically, and NOTIFYING afterwards

INSTALLING updates automatically, WITHOUT notifying afterwards

If you aren't sure, choose your best guess.
(Only software selected in Q6 was shown)

	Checking and Notifying	Checking, Downloading and Notifying	Installing and then Notifying	Installing Without Notifying
Windows operating system	[ ]	[ ]	[ ]	[ ]
Microsoft Office	[ ]	[ ]	[ ]	[ ]
Anti-virus software	[ ]	[ ]	[ ]	[ ]
Virus definitions or data files for your anti-virus software	[ ]	[ ]	[ ]	[ ]
Firewall software	[ ]	[ ]	[ ]	[ ]
Web browser, like Chrome or Firefox	[ ]	[ ]	[ ]	[ ]
Internet security software	[ ]	[ ]	[ ]	[ ]
Anti-spyware software	[ ]	[ ]	[ ]	[ ]
Adobe products, like Adobe Reader or Flash	[ ]	[ ]	[ ]	[ ]
Java	[ ]	[ ]	[ ]	[ ]
Database, like Oracle or Microsoft Access	[ ]	[ ]	[ ]	[ ]
Graphic design, like Photoshop	[ ]	[ ]	[ ]	[ ]
Multimedia, like iTunes, DVD player	[ ]	[ ]	[ ]	[ ]
Games	[ ]	[ ]	[ ]	[ ]
Communication, like Skype, Instant Message	[ ]	[ ]	[ ]	[ ]
Educational software	[ ]	[ ]	[ ]	[ ]

Q12: Thinking about software installed on the laptop you brought with you that CHECKS for updates, NOTIFIES you that an update is ready, but does NOT automatically install it, how long after being notified do you typically install the update?
(Only software selected in Q11 as Checking and Notifying was shown)

	Right Away	Later	Never
Windows operating system	( )	( )	( )
Microsoft Office	( )	( )	( )
Anti-virus software	( )	( )	( )
Virus definitions or data files for your anti-virus software	( )	( )	( )
Firewall software	( )	( )	( )
Web browser, like Chrome or Firefox	( )	( )	( )
Internet security software	( )	( )	( )
Anti-spyware software	( )	( )	( )
Adobe products, like Adobe Reader or Flash	( )	( )	( )
Java	( )	( )	( )
Database, like Oracle or Microsoft Access	( )	( )	( )
Graphic design, like Photoshop	( )	( )	( )
Multimedia, like iTunes, DVD player	( )	( )	( )
Games	( )	( )	( )
Communication, like Skype, Instant Message	( )	( )	( )
Educational software	( )	( )	( )

Q13: Have you ever changed the settings for whether software automatically CHECKS for updates?

( ) Yes

( ) No

( ) I don't know

Q14: Have you ever changed the settings for whether software updates are INSTALLED automatically?

( ) Yes

( ) No

( ) I don't know

Q15: For each of the following types of software you have installed on the laptop you brought with you, how likely would you be to interrupt whatever task you were using the software for, to install a **security update**? Please rate how likely you would be to to do this from Very Unlikely to Very Likely:
(Only software selected in Q6 was shown)

	Very Unlikely	Unlikely	Undecided	Likely	Very Likely
Windows operating system	( )	( )	( )	( )	( )
Microsoft Office	( )	( )	( )	( )	( )
Anti-virus software	( )	( )	( )	( )	( )
Virus definitions or data files for your anti-virus software	( )	( )	( )	( )	( )
Firewall software	( )	( )	( )	( )	( )
Web browser, like Chrome or Firefox	( )	( )	( )	( )	( )
Internet security software	( )	( )	( )	( )	( )
Anti-spyware software	( )	( )	( )	( )	( )
Adobe products, like Adobe Reader or Flash	( )	( )	( )	( )	( )
Java	( )	( )	( )	( )	( )
Database, like Oracle or Microsoft Access	( )	( )	( )	( )	( )
Graphic design, like Photoshop	( )	( )	( )	( )	( )
Multimedia, like iTunes, DVD player	( )	( )	( )	( )	( )
Games	( )	( )	( )	( )	( )
Communication, like Skype, Instant Message	( )	( )	( )	( )	( )
Educational software	( )	( )	( )	( )	( )

Q16: For each of the following types of software you have installed on the laptop you brought with you, how willing would you be to interrupt whatever task you were using the software for, to install **OTHER, NON-security updates**? Please rate how likely you would be to to do this from Very Unlikely to Very Likely:
(Only software selected in Q6 was shown)

	Very Unlikely	Unlikely	Undecided	Likely	Very Likely
Windows operating system	( )	( )	( )	( )	( )
Microsoft Office	( )	( )	( )	( )	( )
Anti-virus software	( )	( )	( )	( )	( )
Virus definitions or data files for your anti-virus software	( )	( )	( )	( )	( )
Firewall software	( )	( )	( )	( )	( )
Web browser, like Chrome or Firefox	( )	( )	( )	( )	( )
Internet security software	( )	( )	( )	( )	( )
Anti-spyware software	( )	( )	( )	( )	( )
Adobe products, like Adobe Reader or Flash	( )	( )	( )	( )	( )
Java	( )	( )	( )	( )	( )
Database, like Oracle or Microsoft Access	( )	( )	( )	( )	( )
Graphic design, like Photoshop	( )	( )	( )	( )	( )
Multimedia, like iTunes, DVD player	( )	( )	( )	( )	( )
Games	( )	( )	( )	( )	( )
Communication, like Skype, Instant Message	( )	( )	( )	( )	( )
Educational software	( )	( )	( )	( )	( )

Q17: Which of these statements do you agree with the most? Please drag-and-drop the statements below to rank them according to your level of agreement with each statement, from (1) Most Agreement to (5) Least Agreement:

1. Installing a software update repairs software (e.g., fixes bugs or malfunctions) and makes my computer more reliable.
2. Installing a software update improves software so that it works better and can do new things.
3. Installing a software update protects software so that it is less vulnerable.
4. Installing a software update is routine maintenance that keeps my computer in good working order.
5. Installing a software update keeps my computer "up to date" so it doesn't fall behind or become obsolete as quickly.

Q18: Was it difficult for you to rank the statements?

( ) No

( ) Yes (Please explain) _____

Q19: How often have you experienced an update that caused your computer to stop working properly?

( ) Never

( ) Rarely

( ) Sometimes

( ) Often

( ) Very Often

Q20: How worried are you about updates causing your computer to stop working properly?

( ) Never thought about this before

( ) Not worried

( ) Slightly worried

( ) Worried

( ) Very worried

Q21: Have you ever had one of the following experiences? Please check all that apply:

[ ] Received a phishing message or other scam email

[ ] Warning in a web browser that says, "This site may harm your computer?"

[ ] Unwanted popup windows

[ ] Computer had a virus

[ ] Someone broke in or "hacked" the computer

[ ] Stranger used your credit card without your knowledge or permission

[ ] Identity theft more serious than use of your credit card number without permission

Q22: How familiar are you with the following Internet-related terms? Please rate your familiarity with each term below from None (no understanding) to Full (full understanding):

	None	Little	Some	Good	Full
RSS	( )	( )	( )	( )	( )
Reload	( )	( )	( )	( )	( )
Widget	( )	( )	( )	( )	( )
Spyware	( )	( )	( )	( )	( )
Proxypod	( )	( )	( )	( )	( )
Tagging	( )	( )	( )	( )	( )
Cache	( )	( )	( )	( )	( )
Frames	( )	( )	( )	( )	( )
Newsgroup	( )	( )	( )	( )	( )
PDF	( )	( )	( )	( )	( )
Torrent	( )	( )	( )	( )	( )
Malware	( )	( )	( )	( )	( )
Wiki	( )	( )	( )	( )	( )
Podcasting	( )	( )	( )	( )	( )
Favorites	( )	( )	( )	( )	( )
Blog	( )	( )	( )	( )	( )

Q23: Have you ever worked in a "high tech" job such as computer programming, IT, or computer networking?

( ) Yes

( ) No

( ) Other (please specify) _____

Q24: How old are you? Please type your answer here: _____

Q25: What is the last grade or class you completed in school?

( ) None, or grades 1-8

( ) High school incomplete (grades 9-11)

( ) High school graduate (grade 12 or GED certificate)

( ) Technical, trade or vocational school AFTER high school

( ) Some college, no 4-year degree (includes associate degree)

( ) College graduate (B.S., B.A., or other 4-year degree)

( ) Post-graduate training/professional school after college (toward a Masters/Ph.D., Law or Medical school)

( ) Post-graduate degree (Masters/Ph.D., Law or Medical school)

( ) I don't know

( ) Other (please specify) _____

Q26: What is your gender?

( ) Man

( ) Woman

( ) Prefer not to answer

Q26: What is your race?

( ) American Indian or Alaska Native

( ) Asian or Pacific Islander

( ) Black or African-American

( ) Hispanic or Latino

( ) White

( ) Other (please specify) _____

Harder to Ignore?

Revisiting Pop-Up Fatigue and Approaches to Prevent It

Cristian Bravo-Lillo
cbravo@cmu.edu

Lorrie Cranor
lorrie@cs.cmu.edu

Saranga Komanduri
sarangak@cmu.edu

Stuart Schechter
stus@microsoft.com

Manya Sleeper
msleeper@cmu.edu

ABSTRACT

At SOUPS 2013, Bravo-Lillo et al. presented an artificial experiment in which they habituated participants to the contents of a pop-up dialog by asking them to respond to it repeatedly, and then measured participants' ability to notice when a text field within the dialog changed. The experimental treatments included various *attractors*: interface elements designed to draw or force users' attention to a text field within the dialog. In all treatments, researchers exposed participants to a large number of repetitions of the dialog before introducing the change that participants were supposed to notice. As a result, Bravo-Lillo et al. could not measure how habituation affects attention, or measure the ability of attractors to counter these effects; they could only compare the performance of attractors under high levels of habituation. We replicate and improve upon Bravo-Lillo et al.'s experiment, adding the low-habituation conditions essential to measure reductions in attention that result from increasing habituation. In the absence of attractors, increasing habituation caused a three-fold decrease in the proportion of participants who responded to the change in the dialog. As with the prior study, a greater proportion of participants responded to the change in the dialog in treatments using attractors that delayed participants' ability to dismiss the dialog. We found that, like the control, increasing habituation reduced the proportion of participants who noticed the change with some attractors. However, for the two attractors that forced the user to interact with the text field containing the change, increasing the level of habituation did not decrease the proportion of participants who responded to the change. These attractors appeared resilient to habituation.

1. INTRODUCTION

Operating systems, browsers, and other software frequently interrupt user workflow with often-irrelevant security warning dialogs. This abundance has been mentioned repeatedly as a problem in usable security research [2, 4, 5, 6, 7, 10, 11]; most authors seem to agree that dialogs are overused, and that when reaching a dialog a high proportion of users will dismiss it because they are already fatigued. Computer users have also self-reported habituation to dialogs. Krol et al. conducted a lab study wherein participants were exposed to two similar pop-up dialogs [8]. 81% of participants clicked through the dialogs; 45% of participants freely mentioned desensitization as a reason for ignoring the dialogs.

Habituation is a simple form of learning in which "repeated or prolonged exposure to a stimulus results in gradual reduction in responding" [9]. After an extensive review, Thompson and Spencer found nine distinctive features of habituation [12]. For example, a) the decrease in response is usually exponential on the number of exposures, b) if the stimulus is taken away, the original response usually reappears in time, c) if repeated series of habituation training and spontaneous recovery are given to a person, habituation becomes progressively faster, d) the weaker the stimulus, the faster and/or stronger habituation becomes (strong stimuli usually show no significant habituation effects), and e) habituation to a given stimulus has been shown to generalize to other stimuli.

In 2013, Akhawe and Felt conducted a large study on telemetry data collected from SSL, malware, and phishing warnings in Chrome and Firefox [1]. In this study, Chrome users were more than twice as likely to ignore SSL warnings as Firefox users. Unlike Firefox, Chrome does not have an exception storing mechanism for certificate errors, and the authors suggest this as one possible reason for the disparity. Chrome users see a warning on each interaction with a self-signed certificate, which could result in many false positives and produce habituation. The authors call this "warning fatigue" and provide timing data that is consistent with this hypothesis [1].

A number of studies have found that browser dialogs resembling those dialogs that participants encounter frequently are more likely to be ignored by participants in laboratory experiments than less-familiar designs [6, 10, 11]. In addition, prior studies have found evidence of habituation beginning to occur after just one or two exposures to a new dialog [2, 4, 11]. However, these studies were not specifically designed to measure the impact of habituation. They did not completely control for other factors that might have been responsible for users' behavior and did not measure the impact of varying levels of habituation.

Bravo-Lillo et al. presented three experiments designed to measure the impact of user-interface modifications created to direct users' attention in security dialogs [3]. In the first two experiments, these interface elements, termed *attractors*, were used in a software installation dialog. The researchers used attractors to direct participants' attention to a *salient field*, a text field that contained information that would allow users to differentiate between harmless and malicious scenarios. For example, the *Swipe* attractor required participants to swipe their mouse over the salient field, which contained the publisher name, to activate the option that presented more risk (i.e., installing software).

In the 2013 study, participants in treatments that used attractors

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

were more likely to be able to differentiate between suspicious and harmless scenarios, and thus appeared more likely to be paying attention to the salient text field. The study's first two experiments used a deceptive ruse of an online game evaluation study to present dialogs to users as if they represented real security decisions with real consequences, so as to maximize ecological validity. However, since participants would not have seen attractors in real-world security dialogs before, it was possible that improvements seen were the result of the novelty of attractors, and that the benefit would wear off over time with the novelty.

To test whether attractors were effective after habituation, the researchers also included a third experiment. Unlike the first two experiments, the researchers did not hide from participants the fact that they were studying interaction with dialogs. Rather, the researchers asked participants to respond to as many dialogs as possible in a five-minute period. The dialog asked participants if they would like to see another dialog, and the only working option was *Yes* (the *habituated* option) with the other option, *No*, having no effect. In this third experiment the researchers used attractors to direct users' attention to a 'Status' field that contained information intended to be irrelevant to their choice of action—the number of dialogs they had dismissed so far. Thus, they attempted to train participants that attractors were only used to direct them to information that was of no use, habituating them to ignore the 'Status' field that the attractors directed them to.

In this third experiment, after a habituation period of either 2.5 minutes or 22 exposures, the researchers placed into the 'Status' field a *habit-breaking* instruction that *was* relevant to the participants' actions. The researchers' new instruction directed participants to choose the now-active *No* option to finish the study early. The researchers measured the performance of each attractor as the proportion of participants who chose this *alternate* option on their first opportunity. They found that, under these habituated conditions, five attractors (*Swipe*, *Type*, *Animated Connector + Swipe*, *Animated Connector + Reveal*, and *Animated Connector + Delay*) performed better than the control [3].

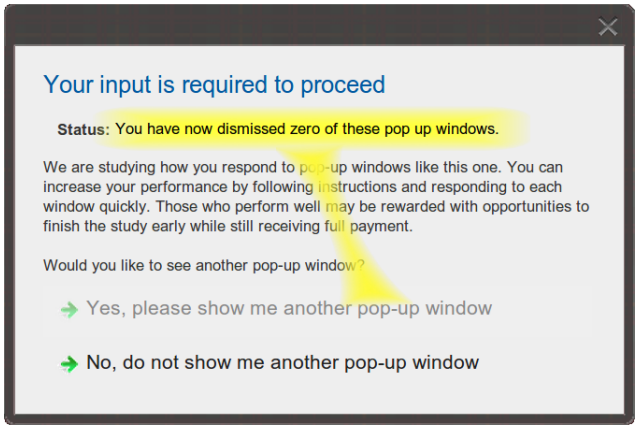
However there was **a serious limitation in this study design.** In Bravo-Lillo et al.'s prior work the researchers included only one level of habituation (changing the dialog box after 22 exposures or 2.5 minutes depending on the type of attractor). Because the researchers did not also include lower habituation conditions, it was not possible to determine whether different attractors were more *resistant* to habituation than the control. Rather, it is possible that some attractors resulted in such large initial increases in attention that they continued to perform well *despite* habituation.

In this paper we expand on Bravo-Lillo et al.'s prior work by replicating and extending the prior experiment for several of their attractors. We overcome the limitations of the previous study by including both *low-* and *high-habituation* conditions. This important extension to the study design allows us to measure attractors' *resistance* to habituation and yields three new findings:

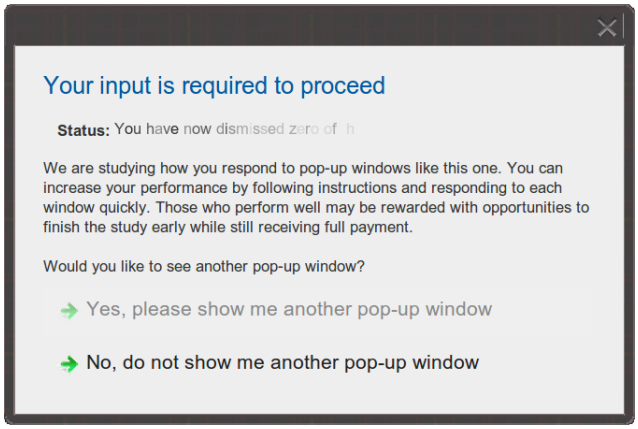
1. **Habituation reduces attention for the baseline experimental task (the control case).** For the control dialog, increasing habituation decreased the proportion of participants who would choose the alternate option (*No*) at the first opportunity by more than a factor of three. This confirms hypotheses about habituation from prior work.

2. **Some attractors failed to show resistance to habituation.** Increasing habituation negatively impacted the performance of some attractors, even though these attractors still outperformed the control in high-habituation conditions. In particular, two attractors that displayed an animation before activating the habituated option (*Yes*) became less effective after habituation. This finding brings new insights to the prior finding from Bravo-Lillo et al. [3] and suggests that the forces of both novelty and habituation might explain the results of the prior study.

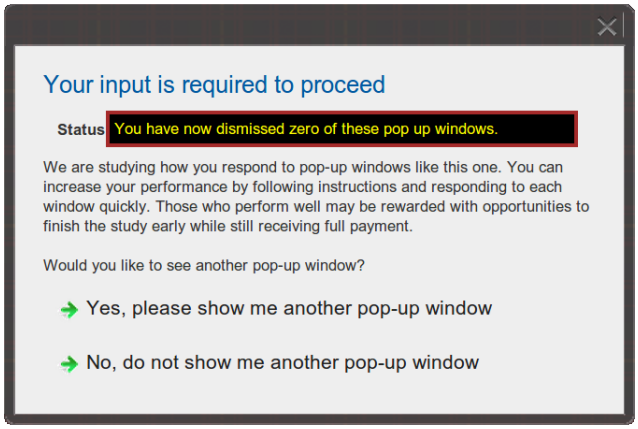
3. **Some attractors showed resistance to habituation.** The length of the habituation period had no measurable impact



(a) Animated Connector + Delay attractor



(b) Reveal attractor



(c) ANSI attractor

Figure 1: Dialogs that are designed to visually draw users' attention to the salient field.

on the performance of two attractors. The first of these two habituation-resistant attractors forced users to swipe their mouse over the salient field before choosing the *Yes* option; whereas the second, more arduous, attractor required them to retype the field's contents. This is the first experimental evidence to demonstrate that some user-interface modifications can significantly reduce, if not entirely prevent, habituation from sapping users' attention to warnings.

As in the prior experiment [3], we also observed that the usability cost of the swipe attractor seems to decrease with time. Once participants grew accustomed to it, they could respond to dialogs containing this attractor within three to five seconds.

2. ATTRACTORS

An attractor is an interface modification designed to draw or force attention to an information field called the *salient field*. The salient field is the part of the dialog that provides the most important information to aid the user's decision.

We implemented five of the attractors presented by Bravo-Lillo et al. [3]. Four are *inhibitive attractors*, which prevent users from making potentially-hazardous choices until after some period of time has passed or a user performs some action. The inhibitive attractors appear only when a user moves the mouse pointer over the button representing the potentially dangerous option. In security dialogs, this *triggering option* is the option that represents a security risk (e.g., installing software). In our study, the triggering option states *"Yes, please show me another pop-up window."* The attractor is not triggered if the user attempts to select the *No* option.

The *Animated Connector + Delay* (AC + Delay, Figure 1a) attractor is a yellow highlight that first appears behind keywords in the triggering option that relate to the salient field. Over a period of two seconds, the highlighted region progresses in the direction of the salient field, and then fills the background of the field. The attractor disables the *Yes* option for five seconds (hence, the delay).

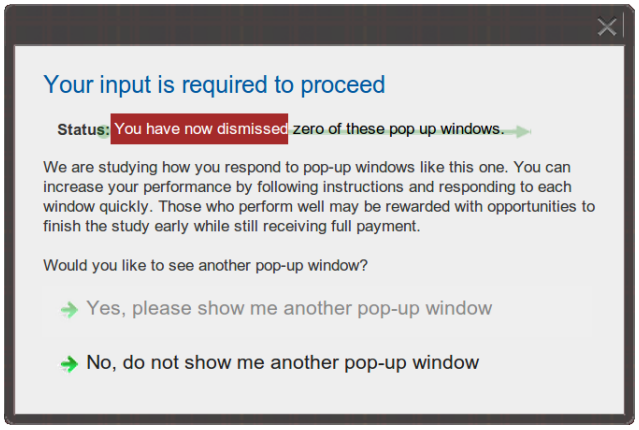
The *Reveal* attractor (Figure 1b) first hides the contents of the salient field, then progressively animates it back in a random fashion, mostly from left to right, over a period of five seconds. The motion and randomization are intended to help users notice each letter as it appears.

The *Swipe* attractor (Figure 2a) disables the *Yes* option until the user moves her mouse from left to right over the salient field. As the mouse moves over each letter, that letter becomes highlighted. If the user moves her mouse over the triggering option before swiping, a pop-up message appears that explains how to swipe and illustrates the swiping motion with an animated cursor.

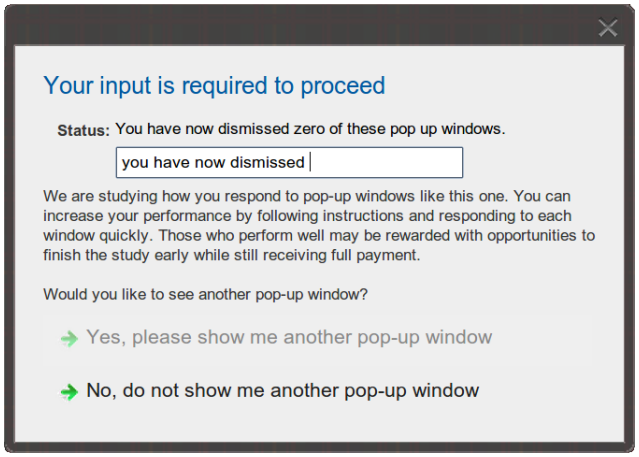
The *Type* attractor (Figure 2b) requires the user to retype the contents of the salient field (no pasting allowed). Bravo-Lillo et al. had included this treatment with the assertion that it would be quite difficult to type text without paying attention to it. We also include this treatment in part to measure effects that may confound our experiment's ability to measure attention (such as frustration with the tedium of a task).

We also included the non-inhibitive *ANSI* attractor from the prior study (Figure 1c), which gives the salient field a black background and high-contrast yellow text to draw attention to it. This treatment helps measure the impact of novel, attention-grabbing stylistic changes that are not accompanied by time delays or required actions.

In Bravo-Lillo et al.'s prior work they also implemented several combined attractors (*Animated Connector + Swipe* and *Animated Connector + Reveal*) [3]. We tested only the uncombined versions in our study.



(a) Swipe attractor.



(b) Type attractor.

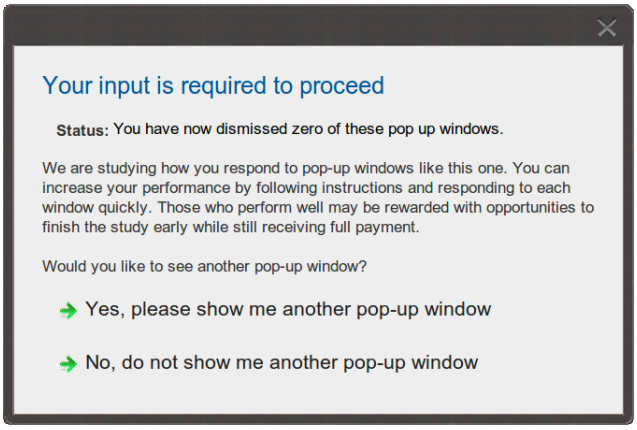
Figure 2: Dialogs that include attractors that require users to interact with the salient field.

3. STUDY DESIGN

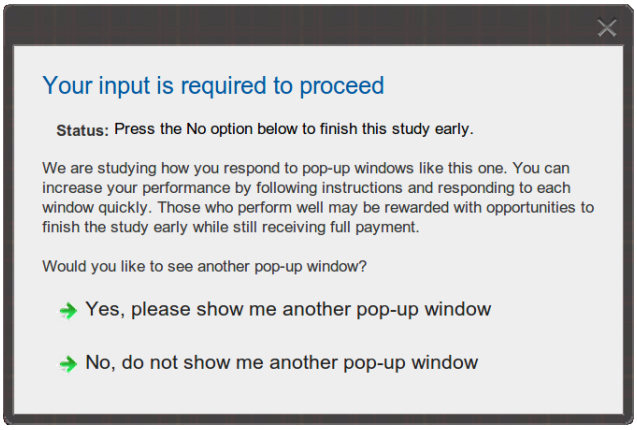
Except where noted, we replicated the experimental methodology documented in Bravo-Lillo et al. [3]. We recruited participants by advertising a human-intelligence task to workers on Amazon's Mechanical Turk, asking them to perform a task in which they would respond to as many dialogs as possible for a fixed time period. We instructed participants that their task was to respond to questions in pop-up windows as quickly as possible over a ten-minute period. We also instructed them to look for opportunities to finish the study early.

During a *habituation period*, we displayed the dialog shown in Figure 3a. In this dialog, the contents of the information field, labeled "Status", alternated between the message "You have now dismissed n of these pop up windows" and "n pop up windows have been dismissed so far," where n was written in words (not digits). This iteration between two irrelevant messages was to ensure that attention to the 'Status' field was due to its content and not to its replacement by another message.

During the habituation period we prevented the *No* option from having any effect, though we did not change its appearance to indicate that we had disabled it. By removing one of the two available options we effectively forced participants to choose the *Yes* option to dismiss the dialog so as to habituate them to clicking this option.



(a) Dialog shown during the habituation phase. The Status field displayed alternatrf between the sentence shown in the dialog, and "N dialogs have been dismissed so far."



(b) Dialog shown during the test phase. Note the change in the Status field.

Figure 3: Control dialogs used in the experiment.

We displayed each dialog at randomly selected coordinates within a participant's browser. If we detected 15 seconds of inactivity we warned participants that we would exclude those who were inactive for 30 seconds or more.

The dialog we used in this experiment differs slightly from prior work in that we used the phrase "pop up windows" to describe the dialogs participants were asked to dismiss, while the prior study asked participants to dismiss "questions." We made this change after piloting to reduce participant confusion. Several participants indicated that they expected actual questions when we asked them about problems encountered during the task.

We followed the habituation period with a *test period* during which we presented the same dialog but with the alternate (*No*) option enabled and the contents of the status field replaced with the instruction, "Press the No option below to finish this study early." Participants who read and understood the habit-breaking instruction in the status field discovered that they should stop choosing the habituated option (*Yes*) and instead choose the alternate option (*No*). We terminated the test period when the participant chose the alternate *No* option or their ten minutes were up.

We then presented an exit survey in which we asked participants to recall the contents of the status field, instructing those with no recollection to type *None*. We paid $1.00 to all participants who completed the experiment.

We designed six treatments: one control treatment with no attractors, and five treatments each with one of the five attractors described in the previous section. For every treatment we created a condition for each of four habituation periods, resulting in 24 (6×4) total conditions. Each participant was assigned to a single condition.

We defined the duration of three habituation periods in terms of the number of habituation dialogs the participant would be exposed to (1, 3, and 20 exposures). These habituation periods lasted for as much time as it took participants to dismiss the dialogs they were exposed to. We did not create a zero-exposure habituation period because participants would have been entirely unfamiliar with the dialog and attractors.

We defined the duration of the fourth habituation period in units of time: 150 seconds, plus whatever additional time was required to dismiss the habituation dialog that was present at the moment the 150-second period expired. This corresponds to the 2.5 minute condition used by Bravo-Lillo et al. For this fixed-time treatment, the number of dialog exposures varies between participants, even within the same condition.

The addition of low-habituation conditions is what most differentiates our study from Bravo-Lillo et al.'s prior work. In the prior study, the researchers tested all participants at one high-habituation point (150 seconds or 22 exposures depending on the attractor). In addition, to accommodate longer habituation periods in our study, we told participants they would be spending ten minutes on our task, instead of the five minutes advertised by Bravo-Lillo et al. (primarily needed for the 20-exposure habituation period with participants in the Type condition).

As in the prior work, a participant is considered "attentive" if he or she chose the *No* option on his/her first test trial—the first trial in which he/she received the habit-breaking instruction.

To analyze the impact of habituation on each attractor, we examine the *habituation odds ratio* between two given habituation conditions. This is the ratio of participants who complied with the habit-breaking instruction in the lower of the two habituation conditions over that in the higher-habituation condition. This yields a $2 \times 2 \times 2$ contingency table: 2 treatments (attractor vs. control) $\times$ 2 habituation conditions (lower vs. higher habituation) $\times$ 2 outcomes (complied with the new instruction or did not). To test the null hypothesis that habituation caused the same reduction in the proportion of participants who complied with the instruction, regardless of treatment, we build a log-linear model without second-order interactions and use a likelihood-ratio test to compare this model to the observed data. If the observed data deviates significantly from the expected model, it indicates that the treatment might have an effect on the habituation odds ratio.

4. RESULTS

We ran this experiment from May 28 until June 09, 2013. We recruited a total of 3,071 participants for the study and 2,567 finished. Participants were 29.4 years old on average (σ=10.1 years), 55% male, 77% Caucasian, and the top two reported occupations were "student" (25%) and "unemployed" (15%). Based on user-agent strings, 60% of participants used Chrome, 37% used Firefox, and 3% used Internet Explorer.

For each participant, we consider the outcome a success if the participant chose the *No* option in response to the first dialog (test trial) in which they were instructed to do so, *complying* with this instruction. The *compliance rate* is the fraction of the participants in each condition who complied. We used a binomial outcome representing the result of the first test trial (complied vs. did not comply) with independent variables for the length of the habituation period

	Fixed exposure count							Fixed exposure time		
	1 exposure		3 exposures		20 exposures			150 seconds		
	med. time	R_{1e} (*No·Yes*)	med. time	R_{3e} (*No·Yes*)	med. time	R_{20e} (*No·Yes*)	$P\left[\frac{R_{1e}}{R_{20e}} = \frac{R^c_{1e}}{R^c_{20e}}\right]$	med. exp.	R_{150s} (*No·Yes*)	$P\left[\frac{R_{1e}}{R_{150s}} = \frac{R^c_{1e}}{R^c_{150s}}\right]$
Control	10 sec	50·56	3.4 sec	43·64	1.2 sec	24·90	—	192	7·99	—
ANSI	10.9 sec	57·55	3.9 sec	49·58	1 sec	15·95	$= 0.1333$	198	13·94	$= 0.3466$
AC + Delay	15.7 sec	89·18	9.8 sec	86·22	6.8 sec	65·43	$= 0.9578$	50	47·60	$= 0.1933$
Reveal	14.2 sec	84·25	8.4 sec	81·22	7 sec	57·47	$= 0.6565$	48	59·47	$= 0.0021$
Swipe	39 sec	61·45	6.9 sec	56·48	3.9 sec	59·48	$= 0.0062$	76.5	65·45	< 0.0001
Type	57.4 sec	79·33	16.6 sec	79·25	12.9 sec	86·13	< 0.0001	24	90·14	< 0.0001

Table 1: For each condition, we present the median of participants' response times to their final habituation dialog (labeled *med. time*). We then present a count of the number of participants who chose the 'no' option on the first test trial (complying with the newly-introduced instruction) followed by the count of those who did not. Together, this is the compliance ratio R. Control group ratios are written R^c. The habituation odds ratio is the low-habituation compliance ratio over the high-habituation compliance ratio. To determine whether habituation had a greater or lesser effect in a treatment than in the control, we attempt to disprove the null hypothesis that their odds ratios are equal.

and the treatment (*attractor* or *control*). We present our results in Table 1 and graph the compliance rate as a function of the number of habituation exposures (log scale) in Figure 4. The level of habituation is measured by the number of exposures to the habituation dialog: 1, 3, and 20 exposures for the first three habituation exposures. The number of exposures varied for the fixed-time-period (150 second) conditions, so we use the median number of dialogs dismissed when plotting this point in Figure 4. Downward slopes represent a reduction in compliance.

For our *Control* dialog, which did not contain an attractor, the compliance rate starts low and declines steeply and steadily as the number of habituation exposures grows. The compliance rates of participants who saw the *ANSI* treatment were not significantly better, and were actually worse (though not significantly so) for the 20-exposure condition.

The two attractors that impose a delay but do not force the user to interact with the salient field, *Animated Connector + Delay* and *Reveal*, did best in the low-habituation conditions, but saw dramatic declines in compliance, with slopes similar to those seen for the control. We use habituation odds ratios to compare reductions in compliance in these attractor treatments with the reduction for the control. We are unable to reject the null hypothesis that the difference in reductions in compliance between the one-habituation-exposure condition and the 20-habituation-exposure condition was due to chance, suggesting that these conditions might not be more resistant to habituation than the control. The likelihood-ratio test yields a probability of the difference occurring under the null hypothesis of $p = 0.6565$ and $p = 0.9578$ (see Table 1).

In contrast, we were surprised to see the compliance rate for *Type* grow with the number of exposures. A possible explanation is that participants' motivation to comply with the instruction, and to end the experiment early, may have increased as they grew tired of the task. This would represent a countervailing force that overpowers the minimal impact of decreased attention for this attractor. This force might be larger if users are particularly annoyed by an attractor. We use the habituation odds ratio to compare the (nonexistent) reduction in compliance due to habituation in the *Type* treatment with the threefold reduction for the control. We reject the null hypothesis that the difference in reductions in compliance between the one-habituation-exposure condition and the 20-habituation-exposure condition was due to chance, as the test yields a probability of the difference occurring under the null hypothesis of $p < 0.0001$. The same is true comparing the one-exposure condition with the 150-second condition (see Table 1).

Following *Type*, the *Swipe* attractor was second most resistant to habituation, with negligible reductions in compliance as habituation increased. Again we use the habituation odds ratio to compare this with the control. We reject the null hypothesis that the difference in reductions was due to chance with $p < 0.0062$ when looking at 1 vs. 20 exposures and $p < 0.0001$ when looking at 1 exposure vs. 150 seconds.

The lower reductions in compliance for *Animated Connector + Delay* and *Reveal* were not statistically significantly better than the control. Yet, the compliance rates for these two delay-inducing attractors were not far from that of *Swipe* under conditions of high habituation. Rather, they start from such a high initial level of compliance that they remain competitive even after significant reductions due to habituation.

However, the relative benefits of *Swipe* become more apparent when we examine the time required to interact with this attractor once participants are familiar with it. We measured the time each participant took to dismiss the last habituation dialog and calculated the median for each condition. Figure 5 shows the 25th, 50th, and 75th percentile dismissal time for the final habituation dialog in each condition. The time spent dismissing a dialog containing no useful information represents one component of the burden that attractors impose on users. Users quickly became efficient at interacting with *Swipe*. After three habituation exposures, participants learned to interact with the *Swipe* attractor as quickly as they did in the delay-based attractors. After 20 exposures, they were nearly twice as efficient in interacting with it as they were in the delay-based attractors. Nearly 75% of participants using the *Swipe* treatment were able to dismiss the 20th habituation dialog within five seconds.

4.1 Limitations

To create an experimental task that would allow us to vary habituation, we opted for a design that was necessarily artificial. Real-world habituation takes place over long periods of time. Security dialogs tend to be viewed one at a time with longer intervals between views, rather than rapidly during a ten-minute period. Also, users might use context in conjunction with the information in security dialogs to make a decision. Thus, users may behave differently when habituated in a more natural setting.

Since different attractors impose different delays, it was not possible to isolate the habituation effects of time and exposure count. Fortunately, this limitation does not appear to impact our conclusions. For the 20-exposure habituation conditions, the *Control* and *ANSI* dialogs required the least amount of time to complete, yielding the shortest habituation time periods, yet they saw the greatest reduction in compliance. In comparison, completing 20 trials took the most time for participants in the *Type* treatment, yet *Type* saw

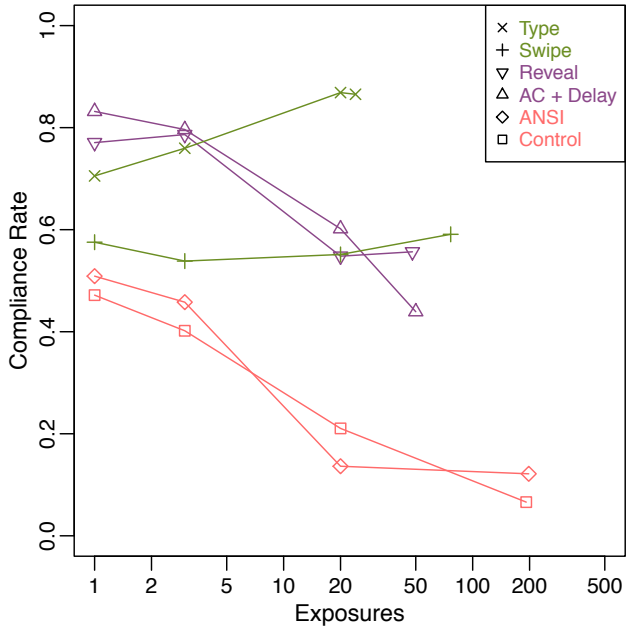


Figure 4: Compliance of participants to the instruction to click *No* in response to the first dialog in which they were asked to do so. The compliance rate is the number of participants who chose *No* over the total number of participants in that condition. The data from which this graph was generated can be found in Table 1.

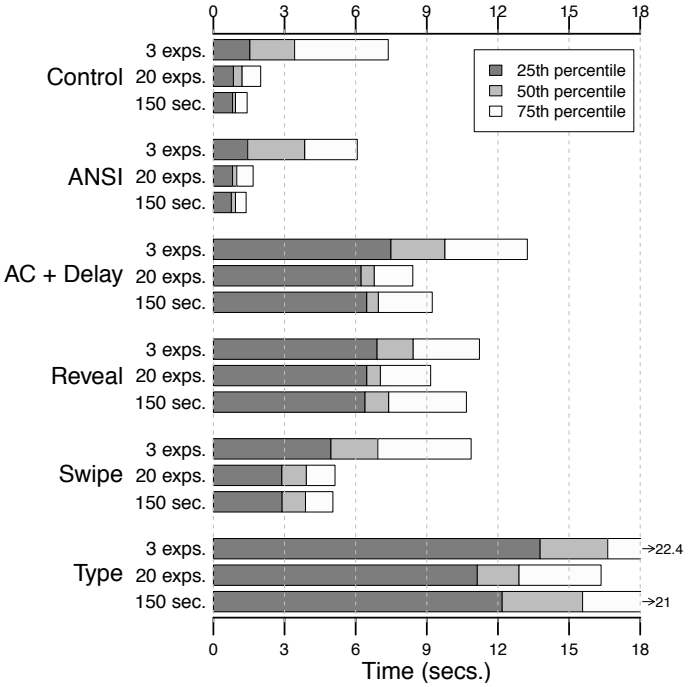


Figure 5: 25th, 50th and 75th percentiles of participants' response time to the last habituation dialog.

an increased rate of compliance due to habituation.

Although at the beginning of the task we instructed participants to look for opportunities to finish the study early, some participants may not have paid attention to that directive or may have felt obligated to keep clicking on *yes* despite receiving the instruction in the 'Status' box. As our experimental design replicates that of Bravo-Lillo et al. [3], the same concerns may apply to their prior experiment.

One way to examine whether participants persevered with the experiment (clicking *yes*) despite having read the instruction to finish is to look to see if participants spent time resolving the conflict between the stated time period of the study and the instruction to finish early. If participants had noticed the change in the 'Status' box, they would presumably require some time to process what they had read. Searching for this decision lag, we examined the 50th, 75th, 90th, and 95th percentile response times at both the last habituation trial and the first test trial. We didn't find any. For example, examining the Control treatment presented with the 150-seconds habituation period, there was only a 10 ms increase in response time from the last habituation trial to the first test trial, at the 50th percentile. We found a decrease in response time at the 75th, 90th, and 95th percentile. If even 5% of participants who chose 'yes' did so after reading the instruction and deciding to ignore it, the 95th percentile should show an increase in response time. The same was true when we examined every control and ANSI treatment. In contrast, when these participants eventually chose 'no' in later trials, the response time increased by a factor of 3 or more, depending on the specific percentile.

We also asked participants who clicked *yes* in the first test trial to explain their behavior. The most popular reason given was not noticing, among myriad other reasons (e.g., one participant reported understanding the instruction and deciding to follow it, but then accidentally clicking *yes* out of sheer habit). The answers reveal only a small minority who believed that we wanted them to persevere or that they would be rewarded for persevering despite our instruction that they should look for opportunities to finish the study early.

One factor that may have affected our participants' responses was the appearance of the *No* button during the habituation period. When we disabled the functionality of the button we did not change its appearance to reflect that it was disabled, lest it would transition from a disabled to enabled appearance at the same time the habit-breaking instruction first appeared. Had the option changed appearance, we would not have been able to separate the effect of participants noticing this change from the effect of participants noticing the habit-breaking instruction. However, it's possible that, as a result of this design choice, some participants disregarded the habit-breaking instruction because they believed clicking the *No* option would have no effect. Fortunately, we see no reason this behavior should be any more likely to occur in one condition than another, and so it should not impact cross-group comparisons.

5. DISCUSSION

While Bravo-Lillo et al.'s prior study demonstrated that some attractors performed well under conditions of heavy habituation, it was not clear whether these attractors were actually resistant to habituation. By testing four levels of exposure we found that some of the attractors that performed well in Bravo-Lillo et al.'s study, specifically *Reveal* and *AC+Delay*, become less effective with repeated exposure. As expected, in the absence of attractors, increasing habituation in the control condition caused a three-fold decrease in performance. On the other hand, two attractors, *Swipe* and *Type*, remained effective even after many exposures. These results have implications for security dialog design and also highlight the value of conducting habituation trials.

Of the attractors we tested, the *Type* attractor performed best, but also imposed the greatest usability burden. While an attractor

of this type may not be realistic for the majority of user environments where users can circumvent systems when frustrated, it may be useful in security-critical environments where lack of habituation should be prioritized over the usability burden. While the *Swipe* attractor had a lower compliance rate than *Type*, it was also resistant to habituation, and it demonstrated a reduction in usability overhead over time as users learned to use it more quickly. Thus, *Swipe* may be a good option in environments where the usability burden associated with *Type* is unacceptable.

This study also demonstrates the value of conducting habituation trials at various levels of exposure. In Bravo-Lillo et al.'s prior habituation experiment, all five inhibitive attractors outperformed the control at the one high-habituation level tested. However, our study demonstrates that some of these high-performing attractors actually become less effective with additional exposures. This suggests that when evaluating a real-world security dialog, it would be useful to test habituation at more typical exposure frequencies as well.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. CNS1116934 and DGE0903659, and by the ARCS Foundation.

6. REFERENCES

[1] D. Akhawe and A. P. Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Proceedings of the 22th USENIX Security Symposium*, 2013.

[2] R. Böhme and S. Köpsell. Trained to Accept?: a Field Experiment on Consent Dialogs. CHI'10, pages 2403–2406. ACM, 2010.

[3] C. Bravo-Lillo, L. Cranor, J. Downs, S. Komanduri, R. Reeder, S. Schechter, and M. Sleeper. Your Attention Please: Designing Security-Decision UIs to make Genuine Risks Harder to Ignore. SOUPS '13, pages 1–12. ACM, 2013.

[4] J. C. Brustoloni and R. V. Salomón. Improving Security Decisions with Polymorphic and Audited Dialogs. SOUPS '07, pages 76–85. ACM, 2007.

[5] L. F. Cranor. A Framework for Reasoning about the Human in the Loop. UPSEC'08, pages 1–15, Berkeley, CA, USA, 2008. USENIX Association.

[6] S. Egelman, L. F. Cranor, and J. Hong. You've been Warned: an Empirical Study of the Effectiveness of Web Browser Phishing Warnings. CHI '08, pages 1065–1074. ACM, 2008.

[7] C. Karlof, J. D. Tygar, and D. Wagner. Conditioned-safe ceremonies and a user study of an application to web authentication. In L. F. Cranor, editor, *SOUPS*, ACM International Conference Proceeding Series. ACM, 2009.

[8] K. Krol, M. Moroz, and M. A. Sasse. Don't work. can't work? Why it's time to rethink security warnings. In *7th international conference on Risk and security of internet and systems (crisis)*, pages 1–8, 2012.

[9] D. L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology*. Worth Publishers, 2009.

[10] A. Sotirakopoulos, K. Hawkey, and K. Beznosov. On the Challenges in Usable Security Lab Studies: Lessons Learned from Replicating a Study on SSL Warnings. SOUPS '11, pages 3:1–3:18. ACM, 2011.

[11] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying Wolf: an Empirical Study of SSL Warning Effectiveness. USENIX '09, 2009.

[12] R. F. Thompson and W. A. Spencer. Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological review*, 73(1):16, 1966.

Your Reputation Precedes You: History, Reputation, and the Chrome Malware Warning

Hazim Almuhimedi
Carnegie Mellon University
hazim@cs.cmu.edu

Adrienne Porter Felt
Robert W. Reeder
Sunny Consolvo
Google, Inc.
felt, rreeder, sconsolvo@google.com

ABSTRACT

Several web browsers, including Google Chrome and Mozilla Firefox, use malware warnings to stop people from visiting infectious websites. However, users can choose to click through (i.e., ignore) these malware warnings. In Google Chrome, users click through a fifth of malware warnings on average. We investigate factors that may contribute to why people ignore such warnings. First, we examine field data to see how browsing history affects click-through rates. We find that users consistently heed warnings about websites that they have not visited before. However, users respond unpredictably to warnings about websites that they have previously visited. On some days, users ignore more than half of warnings about websites they've visited in the past. Next, we present results of an online, survey-based experiment that we ran to gain more insight into the effects of reputation on warning adherence. Participants said that they trusted high-reputation websites more than the warnings; however, their responses suggest that a notable minority of people could be swayed by providing more information. We provide recommendations for warning designers and pose open questions about the design of malware warnings.

1. INTRODUCTION

Modern browsers such as Google Chrome and Mozilla Firefox try to stop users from visiting websites that contain malware. Simply visiting an infectious website can be enough to harm a user's computer, via a drive-by download attack. Instead of loading infectious websites, browsers present users with full-page warnings that explain the threat (Figure 1). Because the malware warning's false positive rate is very low [30], our goal is for no one to ignore the warning. Yet, people click through (i.e., ignore) 7% and 23% of Firefox and Chrome malware warnings respectively [5].

As part of an effort to improve the design of Chrome's malware warning, we investigate factors that may contribute to why people ignore such warnings. One hypothesis is that some users trust familiar websites enough to not believe warnings about the familiar websites, leading them to click through the warnings. In this paper, we test this familiarity hypothesis through (a) an analysis of nearly four million actual Google Chrome warning impressions, and (b) a survey-based controlled experiment conducted with 1,397

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

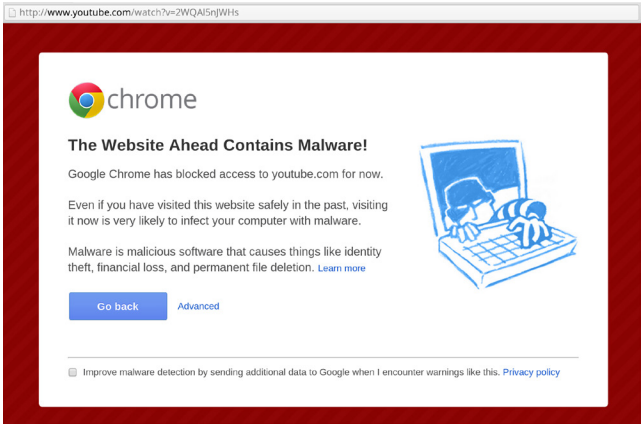


Figure 1: Malware warning in Google Chrome 32

Amazon Mechanical Turk workers. We investigate the impact of people's familiarity with the website they are attempting to visit, as well as how they found out about the website. We also tested minor variations of the instrument used in our survey-based experiment to determine how small wording changes affected responses (e.g., whether or not participants were primed with the word "warning").

Our field data and Mechanical Turk experiment both support our familiarity hypothesis. In our analysis of 3,875,758 malware warning impressions, users were twice as likely to click through the Chrome malware warning if the blocked website was in their browsing history. To further explore why users seem to ignore such warnings, we asked participants in our survey-based experiment about hypothetical warning scenarios. Participants said that it was unlikely that a well-known website would contain malware, so the warning was probably a mistake. They did not appear to realize that even reputable websites can be compromised to temporarily distribute or redirect to malware. However, when participants weren't familiar with the website, they said they would be more likely to play it safe and trust the browser's recommendation.

Contributions. We make the following contributions:

- Through field data and results of an online survey-based experiment, we demonstrate that a person's familiarity with a blocked website has a strong influence on her response to malware warnings.
- We are the first to investigate *why* participants might heed or ignore browser malware warnings.

- Based on the misconceptions and pain points revealed by participants in our survey-based experiment, we provide recommendations for the design of browser malware warnings.
- Through minor variations of our survey instrument, we explore how role-playing, priming, and interactivity affect results of online survey-based warning studies.

1.1 Why Show Malware Warnings?

Ignoring a malware warning carries substantial risk because the false positive rate is very low [30]. This naturally raises a question: why does Chrome let users click through the warning? We could achieve a 0% click-through rate for the warning simply by taking away the ability to proceed.

We don't fully block malicious websites because of the following concerns:

- A determined user might disable the Safe Browsing service to get to the desired content. This would leave the user without protection in the future.
- An unconvinced user could simply open the website in another browser that does not block the website. The user would likely be exposed to the same risk in another browser, but possibly without realizing it.

Thus, our goal is to convince users to heed the warning.

2. BACKGROUND

We explain when and why Google Chrome shows malware warnings. We then cover prior literature on browser warnings, which has primarily focused on SSL warnings.

2.1 Malware Warnings

Google Safe Browsing [3] scans websites for signs of malware or phishing. The service maintains a list of known malware and phishing sites. Google Chrome checks every page load against this list, looking for two things:

1. Is the destination URL on the list?
2. Does the page load resources (e.g., scripts) from third parties that are on the list?

For both conditions, Google Chrome halts the page load and shows a malware or a phishing warning. Users can click on "Advanced" (Figure 1) and then "Proceed at your own risk" (Figure 5) to dismiss the warning and load the page.

The Safe Browsing list includes many websites that primarily function as attack sites. However, legitimate websites can also temporarily end up on the list if they are compromised [30]. Attackers can subvert legitimate websites via vulnerabilities, user-contributed content, advertisements, or third-party widgets [31]. Websites are removed from the list when they no longer pose a risk.

2.2 Related Work

Malware warnings. Microsoft reported that the CTR for Internet Explorer's SmartScreen malware warning was under 5% in 2011 [21]. Akhawe and Felt reported telemetry data from Google Chrome and Mozilla Firefox for malware, phishing, and SSL warnings [5]. Based on their analysis, malware warning CTRs fluctuate in Google Chrome but not in Mozilla Firefox. They did not investigate the degree of fluctuation or its causes. In this paper, we delve further into the fluctuation issue with additional field data and an online, survey-based experiment.

Others have studied users' perceptions of malware in general, without focusing on warnings. Solic and Ilakovac asked electrical engineers and medical personnel about their security habits; all but one participant were concerned enough about malware to use security software [34]. Asgharpour et al. ran a card-sorting exercise to see whether expert and non-expert computer users had similar mental models of malware-related terms [6]. They found that physical world (e.g., locks) and criminal mental models were the best security metaphors for communicating risk to non-experts.

Phishing warnings. Egelman et al. studied phishing warnings and published several recommendations for warning design, including using interruptive (active) warnings and preventing habituation [10]. Egelman and Schechter [11] found that warnings that explain specific threats may reduce clickthroughs compared with warnings that have vague messaging such as "this website has been reported to be unsafe."

SSL warnings. SSL warnings serve a similar purpose: the browser stops a page load, warns the user of risk, and asks the user to make a decision. However, the threat model differs. With an SSL warning, the attacker is on the network; with a malware warning, the attacker is on the destination website. Furthermore, SSL warnings are commonly false positives whereas malware warnings are rarely unwarranted. Thus, it is not clear whether all of the lessons learned from SSL warnings also apply to malware warnings.

Dhamija et al. exposed laboratory study participants to Mozilla Firefox's SSL warnings during simulated phishing attacks [9]. Of their twenty-two participants, only one was able to correctly describe the contents of the warning to researchers. This study demonstrated that people may not pay attention to or understand SSL warnings.

Schechter et al. studied Internet Explorer 7's SSL warning [32]. In their experiment, participants saw SSL warnings while trying to perform tasks on a banking website. The researchers created three versions of the task in which participants used their own credentials, played a role with fake credentials, or played a role with fake credentials and priming. They found a statistically significant difference between the role-playing participants and the non-role-playing participants, but priming had little effect. We follow their lead and similarly test multiple variants of the instrument used in our online survey-based experiment.

Sunshine et al. tested several SSL warnings in an online survey and laboratory study [37]. In their experiment, participants saw warnings on either a banking website or a university library website. Their participants clicked through the SSL warnings at a slightly higher rate for the university library website than for the banking website. We similarly explore the relationship between the website blocked by a warning and participants' willingness to ignore the warning. However, trust plays different roles in SSL and malware warnings. With an SSL warning, the user must evaluate (1) how much she trusts the network connection, and (2) how sensitive the information on the destination website is. With a malware warning, the user must evaluate whether she thinks a website is going to infect her computer.

Sotirakopoulos et al. replicated Sunshine's prior labora-

tory study [35, 37]. Their primary finding was that the laboratory environment had influenced some participants' decisions. For this reason, we do not believe that participants' CTRs in our online survey-based experiment are indicative of their real world CTRs. When intepreting our survey-based experiment's results, we instead focus on differences between scenarios and understanding users' mental models.

Akhawe and Felt showed that Mozilla Firefox's SSL warning has a lower CTR than Google Chrome's SSL warning [5]. In follow-up work, Felt et al. ran a field study to test factors that could explain the difference between the two browsers' warnings [13]. When they ran Firefox's SSL warning in Chrome, it yielded a lower CTR than the default Chrome SSL warning. They found that the imagery, number of clicks, and styling were not responsible for the difference. However, the Firefox-UI-in-Chrome CTR was still higher than the Firefox-UI-in-Firefox CTR. They concluded that demographic factors or other unknown variables besides the warning UI must be influencing user behavior across browsers.

Credibility and trust online. As warning designers, we need users to trust our malware warning more than the infectious target website. To understand users' behavior and trust decisions, we turn to credibility and trust literature.

Fogg et al. identified seven factors that increase or decrease the credibility of websites [15]. Of those factors, five boost website credibility: real-world feel, ease of use, expertise, trustworthiness, and tailoring. Two hurt the credibility of websites: commercial implication and amateurism. In a follow-up study, Fogg et al. asked participants to comment on different aspects of credibility. The most frequently mentioned factor was the "look and feel" of websites. The second most mentioned factor was how well the website was structured. The authors proposed that "Prominence-Interpretation Theory" explains how users evaluate the credibility of a website. First, a user needs to notice an element of the website that increases or decreases its cerdibility. Second, the user needs to decide whether the element increases or decreases the website's credibility [17].

Briggs et al. introduced a "two-process" model of trust: a first impression, followed by careful analysis [7]. They conducted two studies to explore these processes. In the first study, they recruited fifteen participants to participate in sessions about house-purchasing advice. A qualitative analysis of these sessions suggested that the "look and feel" of the website influences the first impression. However, other factors played an important role when participants turned to a more detailed evaluation. To explore these factors, the authors conducted an online survey with more than 2500 participants who sought advised online. The authors identified three factors that influence the detailed evaluation of online advice: source credibility, personalization, and predictability. Further analysis showed that source credibility was the most important factor when users turn to detailed evaluation about online advices.

Kim and Moon conducted four consecutive studies to explore how to trigger a feeling of trust in cyber-banking systems (text-based, videotex, and online interfaces) [23]. They found correlations between design factors and four emotional factors: symmetry, trustworthiness, awkwardness, and elegance. Trustworthiness, in particular, was determined by the main clipart and the color of the interface.

Date	CTR	N	Date	CTR	N
Tu Oct 01	15%	97,585	Tu Oct 15	16%	73,370
We Oct 02	15%	96,076	We Oct 16	18%	85,266
Th Oct 03	15%	104,075	Th Oct 17	15%	68,947
Fr Oct 04	16%	84,165	Fr Oct 18	11%	132,410
Sa Oct 05	15%	80,433	Sa Oct 19	10%	99,778
Su Oct 06	15%	77,931	Su Oct 20	12%	95,163
Mo Oct 07	16%	80,640	Mo Oct 21	14%	91,651
Tu Oct 08	17%	90,356	Tu Oct 22	21%	131,700
We Oct 09	21%	145,893	We Oct 23	18%	121,944
Th Oct 10	21%	96,159	Th Oct 24	24%	151,387
Fr Oct 11	23%	93,059	Fr Oct 25	27%	117,002
Sa Oct 12	15%	79,295	Sa Oct 26	14%	64,740
Su Oct 13	15%	79,134	Su Oct 27	14%	70,713
Mo Oct 14	18%	89,180	Mo Oct 28	15%	59,567

Table 1: Chrome malware warning click-through rates (CTRs) and sample sizes for October 2013. Darker shaded values indicate higher CTRs. Note the wide variance in daily CTRs.

3. FIELD DATA: BROWSING HISTORY

Google Chrome's opt-in statistical reporting allows us to measure how end users respond to malware warnings in the field. This data allows us to see trends in how Chrome users react to malware warnings. We focus on the role of browsing history in users' malware warning decisions.

3.1 Motivation

Users respond very differently to malware warnings depending on the date. Within the last year (2013-2014), we have observed days where the CTR is as low as 7% or as high as 37%. This is a sizable range, and the degree of fluctuation is unique in Chrome: Chrome's SSL and phishing warning CTRs are stable over time [5].

To illustrate this phenomenon, Table 1 depicts the daily variation of the malware warning CTR in October 2013. Although the average is 17%, the daily CTR ranges from 10% to 27% within the month. "High" and "low" days tend to clump together in a series of similar days. The variation is not due to the day of the week.

As shown in Table 1, the CTR noticeably increased during October 22-25, 2013. We looked for changes in the Safe Browsing list that match these dates. Several high- and medium- reputation sites were added to and removed from the Safe Browsing malware list over a few days: `desitvforum.net` (3229 on the Alexa global ranking, 669 on the Alexa India ranking [1]), `php.net` (228 on the Alexa global ranking [2]), and `warriorforum.com` (117 on the Alexa global ranking [4]). This suggested to us that users might react differently to warnings on popular websites.

However, we are also aware of a counterexample on February 9, 2013. The compromise of an advertising network led to malware warnings on popular websites such as ABC News and YouTube. A few news outlets reported the incident, said that the cause was unclear, and recommended that users heed the warning [25, 36]. Social media posts rose to the top of search results, confirming that many people were seeing the warnings.[1] During these events, the daily CTR dropped

[1]For example:
`http://www.zyngaplayerforums.com/showthread.php?1748942-us-bernerverein-ch-malware-warning`,

from 15% to 8%. When the warnings were removed from the popular websites, the CTR returned to 15%. This indicates that the issue might be more complex than popularity alone – news media, word of mouth, and other factors might influence user behavior.

3.2 Hypotheses

The malware warning CTR varies daily, but so does the Safe Browsing list. Could changes in the Safe Browsing list be responsible for how people are reacting to the warning? We hypothesized that:

- H_1: People are more likely to ignore warnings about websites that they have visited before.
- H_2: When popular websites are on the Safe Browsing list, the CTR will be higher. That is, we expect to see a positive correlation between the CTR and the fraction of blocked websites that were previously visited.

3.3 Methodology

We leveraged Google Chrome's opt-in metrics to test our hypotheses. Google Chrome generates statistical reports on how many people click through the malware warning. We extended these reports for Google Chrome 32.

Implementation. We modified the malware warning to query the history manager. The history manager responds with the number of times that the user has previously visited the website that the warning is for. The malware warning then records two separate metrics: the overall CTR, and the CTR specifically for websites that the user has never visited. Only the history status and decision are recorded; the URL itself is *not* included in the statistics.

Sample. We analyzed metrics from the Google Chrome 32 stable channel from January 28, 2014 to February 24, 2014. Our overall sample size is 3,875,758 warning impressions.

Participation. During installation, Chrome users are asked whether they would like to send "crash reports and statistics" to Google. For users who choose to participate, Google Chrome sends statistical reports to Google. The reports are pseudonymous and cannot be traced back to the sending client once they are stored. We added new histogram entries to these reports. The reports do not contain any personal information (e.g., URLs are not allowed in the reports).

Limitations. Browsing history is an imperfect measure of prior experience with a website. Users clear their history and use multiple devices without syncing their history. In these cases, the user's decision will be misattributed to the "new site" distribution instead of the "visited site" distribution.

The "new site" and "visited site" distributions might contain multiple impressions from the same users, both within and across distributions. We rely on our very large sample size to mitigate this source of potential bias.

3.4 Results

Over the 28-day time period, users were twice as likely to ignore warnings about websites that were already in their browsing history. The average CTR for previously visited websites was 25.4%, whereas the average CTR for new websites was 13.1%. Our evidence supports H_1: the difference between the two average CTRs is statistically significant ($p < 0.0001$, one-tailed Z-test of proportions).

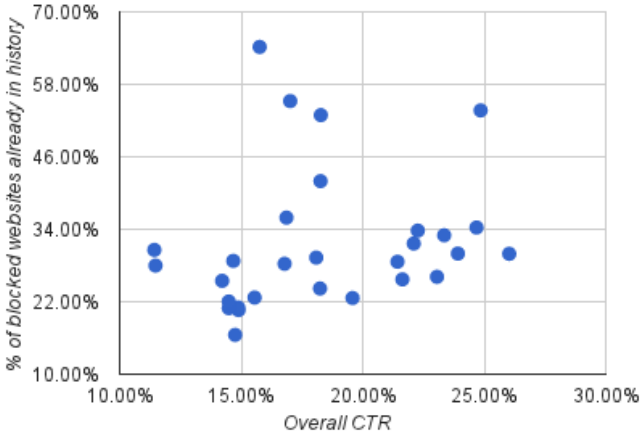


Figure 2: The relationship between the CTR and percentage of blocked websites that were already in the user's browsing history. Each point is a day. For 28 days in January-February 2014.

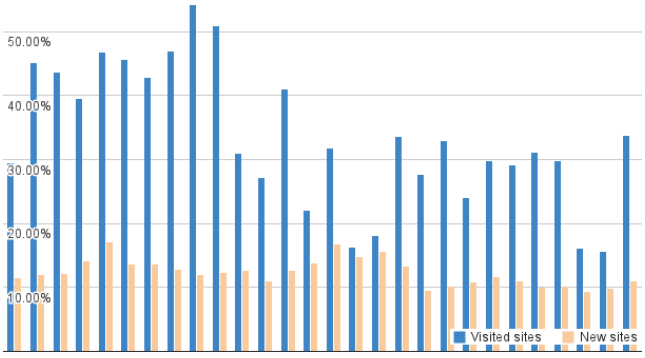


Figure 3: Daily CTR, separated by whether the website was already in the user's browsing history. For 28 days in January-February 2014.

However, the daily CTR is not correlated with the fraction of blocked websites that were previously visited. Figure 2 shows the lack of positive correlation. This means that the number of previously visited websites on the Safe Browsing list is not the cause of the daily variance. A linear regression gave a slope (0.07495) that was not significantly different from 0 ($t = 1.069$, $p = 0.295$), so we fail to reject the null hypothesis for H_2. This is a surprising result: we had expected that H_2 would follow from H_1.

Figure 3 illustrates why our data supports H_1 but not H_2. The CTR for warnings on new websites remains fairly stable over time (9.3% to 17.2%; stdev=2.1%), but the CTR for warnings on previously visited websites varies quite widely (15.6% to 54.3%; stdev=10.9%). Most of the daily variance in the overall CTR can be attributed to the variance within the visited website warnings. This suggests that a second unknown factor — such as reports from the media, word of mouth, or the quality or desirability of the website — may also be influencing user behavior. The unknown factor has a greater effect on user decisions when the destination website is already in the user's browsing history.

`http://answers.yahoo.com/question/index?qid=20130209134718AAhnNZX`, `http://www.reddit.com/r/Malware/comments/187of3/malware_warning_popping_up_everywhere_today/`

4. MTURK: METHODOLOGY

Section 3 showed that users are more likely to ignore a warning if they have visited the destination website before. We hypothesize that this is because prior positive experiences contribute to a website's reputation, and users are less likely to believe malware warnings for high-reputation websites. To explore the role of reputation in malware warning decisions, we set up an online, survey-based experiment.

We asked 1,397 Mechanical Turk workers to tell us how they would react to screenshots of Google Chrome malware warnings. In one experiment, we asked participants to respond to warnings on high- and low-reputation websites (YouTube or an unknown personal blog). In another experiment, we asked participants to respond to warnings that were linked from high- and low-reputation referrers (a friend's Facebook status or a low-quality lyrics website). We also tested minor variations of both experiments to evaluate how the specific question wording affected responses.

4.1 Research Questions

We focus on two questions related to reputation:

- Does the reputation of the referrer (i.e., the source that linked to the warning) affect how users respond to malware warnings?
- Does the reputation of the destination (i.e., the site with the warning) affect how users respond to malware warnings?

Reputation refers to a perception of quality. It can be established via prior personal experience (i.e., browsing history), brand recognition, word of mouth, or other factors.

4.2 Experiment Scenarios

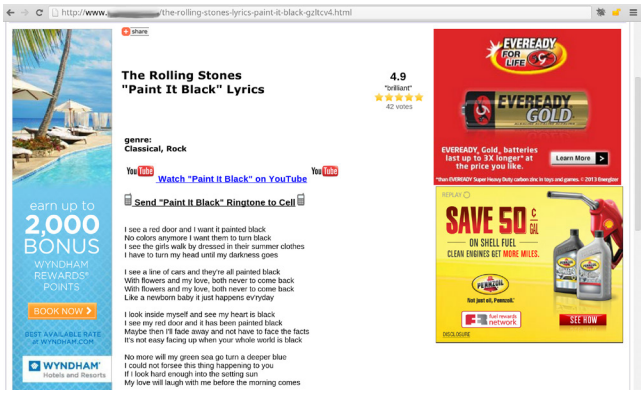
We presented participants with scenarios in which a referrer links them to a destination website with a warning. We created three scenarios (Figure 4):

1. Low-reputation referrer (lyrics website) → high-reputation destination (YouTube)
2. High-reputation referrer (friend's Facebook status) → high-reputation destination (YouTube)
3. High-reputation referrer (friend's Facebook status) → low-reputation destination (low-reputation blog)

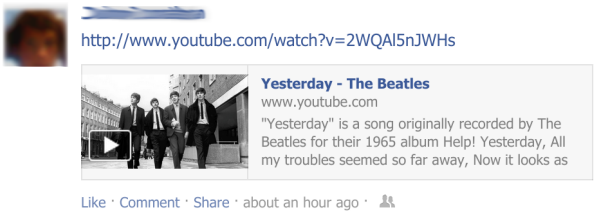
We ran two within-subjects experiments with these scenarios. The **referrer experiment** asked participants about scenarios 1 and 2 in a random order to evaluate the effect of the referrer's reputation. The **destination experiment** asked participants about scenarios 2 and 3 in a random order to evaluate the effect of the destination's reputation.

For the referrer experiment, we chose a friend's Facebook status to represent a high-reputation referrer because Facebook is a common way of exchanging links. We used a lyrics website for the low-reputation referrer because lyrics websites have poor reputations as sources of malware and unwanted advertisements [26, 38].

For the destination experiment, we chose YouTube as an example of a high-reputation destination because it is a highly popular, family-friendly website. We selected a little-trafficked personal blog to represent a low-reputation destination. A branding question at the beginning of the survey



(a) Low-reputation referrer that links to a high-reputation destination



(b) High-reputation referrer that links to a high-reputation destination

(c) High-reputation referrer that links to a low-reputation destination; its URL has been partly obscured for the paper

Figure 4: Screenshots from the scenarios used for the experiments. Each was followed by a screenshot of a Chrome malware warning.

confirmed that participants were familiar with YouTube but not the blog (100% and 0.01% of participants said they were familiar with the two websites, respectively).

We could have also tested a fourth scenario with a low-reputation referrer and a low-reputation destination. However, a pilot study suggested that this was not necessary because participants' self-reported click-through rates for scenario 3 were already close to 0%. As a result, we did not think that a fourth scenario would yield additional results; we decided to focus on the other three scenarios to increase our sample sizes within our budget.

4.3 Wording Choices

Our survey wording could influence the results. To account for this, we tested multiple versions of the two experiments. Prior work has similarly run multiple versions of experiments to look for biases [24,32]. We tested five between-subjects versions of the destination experiment (three roles, priming, and interactive) and three between-subjects versions of the referrer experiment (three roles).

Roles. We asked participants to imagine the scenario as if they were personally experiencing the situation, advising a friend, or pretending to be someone else.

- Personal experience is a natural way to frame a scenario, e.g., *"Imagine that you are visiting..."*
- Researchers use the "helping a friend" role to reduce social desirability bias [14]. We asked participants to help their best friend decide what to do about a warning. For example, *"Imagine that your best friend is visiting www.facebook.com to check friends' latest updates."*
- We asked participants to pretend to be someone else who is completing a task. Researchers use this type of role to reduce the risk of an experiment, reduce social desirability bias, and/or motivate participants to complete an imaginary task [32, 33, 41]. Having a task to complete is intended to mimic real life situations. For example, *"Imagine that you are a moderator of a 'Music Video of the Day' Facebook group that only your friends can join. Your friends post YouTube videos they like to the group, and you visit them to record the number of views. The winner of the day is the most viewed video. Imagine that you are visiting www.facebook.com to check the videos posted to the group today."*

Priming. Prior work offers conflicting guidance on the effects of priming on security research [12, 32, 40]. Thus, we took care to avoid mentioning risk and used neutral language (e.g., "page" or "red page" instead of warning) in all but one version of the experimental survey. One survey variation intentionally began with a paragraph that discussed malicious software and potential risks in order to prime participants. It also used the word "warning" in the prompts.

Interactivity. In one variant of the destination experiment, we provided participants with the ability to read more information about the warnings before deciding. This variant was interactive: participants could choose from any of the available buttons and walk through a series of screenshots until reaching a decision. For example, a participant could select the option "click on 'Advanced' link" to see additional options (Figure 5 shows the additional options). From there, the participant could choose "click on 'Details about the problems on this website'" to see the diagnostic page (Figure 7). This increased the length and complexity of the survey but allowed us to study the effect of providing all of the available options.

4.4 Survey Walkthrough

We created eight surveys: five variations of the destination experiment, and three variations of the referrer experiment. All of the surveys were similarly structured, although the variants had slightly different wording for the scenarios. Each survey had two scenarios. The following illustrates the survey's outline (with a full example in the appendix):

1. Brand familiarity. "Which of these websites have you heard of?" We alphabetically listed the three websites that appear in the survey (Facebook, the blog, and YouTube) and four decoy websites.

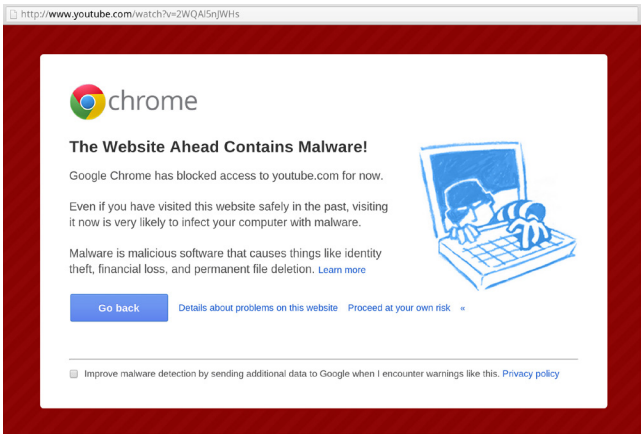


Figure 5: The malware warning, with the "Advanced" options exposed

2. Scenario introduction. For example, "Imagine that you are searching for the lyrics for the song 'Paint It Black'. You find the lyrics on the website shown below. [screenshot]" We then asked a comprehension question to ensure that participants looked at the screenshot. For example, "Which band recorded the song shown in the screenshot above?"
3. Reaction to warning. The survey instructed the participant to imagine she had clicked on a link. It then displayed a screenshot of a Chrome malware warning and asked: "What would you do?" (multiple choice) and "Why?" (short essay).
4. Second scenario. Steps 2 and 3 were repeated for a different scenario.
5. Ramifications. Questions about the ramifications of clicking through the warnings, e.g., "Which outcome is the most likely if you clicked through the red page to proceed from the lyrics website to youtube.com?"
6. Real world behavior. "How would you typically react if you saw a similar red page when trying to visit a website in your day-to-day life?" Also, "Before this survey, had you ever seen a similar red page when trying to visit any website?" If yes: "What happened the last time you saw a similar red page when trying to visit a website?"
7. Demographics. Demographic questions to measure their reputation with the websites in the survey, technical ability, and security knowledge. Also basic demographic information such as age and education level.

The questions were a mix of closed- and open-ended questions, giving us a mix of quantitative and qualitative data.

We randomly assigned participants to one of eight versions of the survey and randomized the order in which the scenarios were displayed. We also randomized the choice order for multiple-choice questions, with a caveat: we kept the choice order constant between similar questions to avoid confusion. (For example, if "Go back" were the first choice for the first scenario's warning question, it would also be the first choice for the second scenario's warning question).

4.5 Recruitment

We used Amazon Mechanical Turk to recruit participants. We posted a task that invited Mechanical Turk workers to "take a survey about how you browse the web." Participants were compensated $1 each for a survey that took 6 minutes on average to complete. We limited the survey to people aged 18 and older in the United States. Participants also had to have an approval rate of 95% or higher. In the instructions, we said that the survey was limited to Google Chrome users. To enforce this constraint, we discarded responses from users who said that they do not use Google Chrome; however, we still paid non-Chrome users to avoid incentivising lying.

We discarded responses from participants who appeared to be cheating. For example, we excluded participants who responded more than once or tried to use incorrect survey completion codes. Each survey also contained two scenario comprehension questions with gold standard answers (see Pages 3 and 4 in the appendix).

4.6 Demographics

We had a total of 1,386 survey responses after excluding submissions that did not meet our requirements. Table 2 shows a summary of participants' demographics. A majority of participants are active Facebook and YouTube users who reported checking Facebook and watching a YouTube video at least once in the week prior to the survey.

Our sample population is likely more tech-savvy than the average Internet user. To assess participants' technical abilities, we asked a multiple-choice question: *"What would you do if your wireless router at home were not working?"* 73% of participants reported that they would fix the problem themselves, which we assume is indicative of relatively high technical confidence. We also asked participants two multiple choice security questions: *"What is a computer firewall?"* and *"What is a public key certificate?"* 44% of participants answered both security questions correctly.

We also asked participants about their highest level of education. About 10% of participants have a post graduate degree, 35% have a bachelor's degree, 31% have some college, 12% have an associates degree, and 11% have a high school diploma or GED. The rest have some high school education without obtaining a diploma.

4.7 Statistical Analysis

We used logistic regression to test for statistical significance of our experimental treatments (destination, referrer, and wording variants). We fitted two logistic regression models, one for the destination experiment and one for the referrer experiment. Except where otherwise noted, p-values for significance testing come from Wald tests at the $\alpha = 0.05$ level of whether the fitted regression coefficients are significantly different from zero. Logistic regression is similar to ANOVA analysis in that it automatically accounts for multiple statistical tests, but unlike standard ANOVA, allows us to model experiments with a binary outcome (in our case, the binary outcome is whether the participant would click through the warning or not).

5. MTURK: LIMITATIONS

Our results must be viewed within the context of the limitations of this type of study.

5.1 Generalizability

Our demographic questions show that most participants are active Internet users who would feel comfortable tinkering with a wireless home router. As such, caution should be exercised in generalizing our results, especially to others with lower levels of Internet exposure. However, our survey population represents an important demographic because active web browsing increases the chances of seeing a warning. Future work could extend this research to groups of users who use the Internet less and are less comfortable with technology.

5.2 Interpretation of Study CTRs

Our experiment asked participants how they would react to warnings under hypothetical circumstances. These artificial conditions differ from real life; our online tasks lacked the urgency that participants might experience in real life, and our experiment posed no real risk. To distinguish our experimental survey results from field data, we refer to the rate at which participants say they would proceed through a warning as the *self-reported click-through rate* (SRCTR).

The primary goal of our work is to evaluate how the reputations of referrers and destinations influence behavior. To this end, we compare SRCTRs between high- and low-reputation conditions. Any bias inherent in our study methodology applies equally to the different conditions, so participants were not biased in favor of any particular condition. In addition, the effect of any inherent bias is minimized by randomizing the order of the tasks (e.g. low-reputation task first), the careful wording of the survey (e.g. using different roles, priming vs. no priming), and the random assignment of participants to different conditions. We therefore interpret SRCTRs as being able to reveal differences between conditions even though they may not be indicative of the absolute value of real-world CTRs.

Despite these limitations, we consider participants' statements to be representative of thoughts that would occur in real encounters with malware warnings, even though they might ultimately act differently due to competing priorities.

5.3 Wording Choices

We were concerned that the wording of our survey instrument would introduce bias. To try to account for this, we tested multiple versions of the survey instrument. Table 3 breaks down the results pertaining to the different survey instrument versions by condition and variation.

Roles. The role did not change most participants' responses. We do not observe a significant difference in SRCTR between roles for the high-reputation destination (37%, 38%, 36%), high-reputation referrer (31%, 31%, 33%), or low-reputation referrer conditions (27%, 28%, 24%). However, playing someone else leads to a higher SRCTR for the low-reputation destination condition (3%, 3%, 8%). The difference is small but statistically significant (p=0.04).

Priming. The type of priming that we used did not influence participants' decisions. The "priming" and "personal" variants are identical except for the presence or absence of priming text and the use of the word "warning" in the prompts instead of the word "red page". The priming variant yields a slightly lower SRCTR (31% vs. 37%) for the high-reputation destination condition, but the difference is not statistically significant (p=0.28). For the low-reputation

Table 2: Characteristics of online, survey-based experiment participants

Experiment	Word Variant	N	% Male	Mean Age	Tech Confident	Security Savvy	% actively use... Facebook	YouTube
Destination exp	Personal ("you")	174	58%	30	76%	42%	87%	96%
Destination exp	Helping a friend	174	54%	30	67%	38%	82%	94%
Destination exp	Playing someone else	173	62%	30	72%	42%	83%	94%
Destination exp	Priming + personal	175	59%	32	74%	59%	86%	94%
Destination exp	Interactive + personal	174	59%	33	75%	47%	87%	93%
Referrer exp	Personal ("you")	172	54%	31	73%	49%	89%	96%
Referrer exp	Helping a friend	171	56%	31	72%	41%	88%	98%
Referrer exp	Playing someone else	173	67%	31	75%	46%	79%	93%

Table 3: Results for the online, survey-based experiment. Darker shaded values indicate higher SRCTRs.

Experiment	Wording Variant	High-Reputation SRCTR	N	Low-Reputation SRCTR	N	Aggregate SRCTR
Destination experiment	Personal ("you")	37%	159	3%	171	19%
Destination experiment	Helping a friend	38%	158	3%	160	20%
Destination experiment	Playing someone else	36%	151	8%	156	22%
Destination experiment	Priming + personal	31%	158	4%	169	17%
Destination experiment	Interactive + personal	15%	162	2%	167	9%
Referrer experiment	Personal ("you")	31%	163	27%	161	29%
Referrer experiment	Helping a friend	31%	166	28%	165	30%
Referrer experiment	Playing someone else	33%	162	24%	161	28%

destination condition, the two SRCTRs are within a percent. This suggests that the type of priming that we used has little effect on participants' responses. This finding is similar to some prior findings about priming in security studies [12,32], although it conflicts with others [40].

Interactivity. For the high-reputation destination, participants in the interactive variant were less likely to proceed than participants in the non-interactive "personal" variant (15% vs. 37%). The difference is statistically significant ($p<0.0001$). The difference is most likely explained by an extra step that participants in the interactive variant had to take in order to click through the warning: they had to first choose an "Advanced" option, while those in non-interactive conditions had the option to click through on the first screenshot they saw.

6. MTURK: RESULTS

We present the results of our experiments in terms of self-reported click-through rates (SRCTRs) and participant quotes. We also present common misconceptions and points of frustration from the short essay responses.

6.1 Destination Reputation

We asked participants to respond to warnings on high- and low-reputation sites, and we find that the destination's reputation strongly affects how participants react to hypothetical warning scenarios. As Table 3 shows, many more participants claim that they would ignore the warning for a high-reputation destination and heed a warning for a low-reputation destination. The difference between the two SRCTRs is statistically significant overall ($p<0.0001$).

Many participants discussed brand reputation and prior personal experience. E.g.,

> *I have never heard of this site [the blog] so I wouldn't trust it.*
>
> *YouTube is well known website. I'd assume that the malware block is in error.*
>
> *Because I frequent youtube.com a lot and I have never gotten any malware*
>
> *Youtube.com is a trusted site that I use almost everyday and have not had any problems with.*

A small number of participants also noticed that the blog is hosted on Blogspot. They said that they would proceed to the blog because they trusted Blogspot.

Additionally, there was a correlation between the reputation of the destination and participants' perceived risk of ignoring the warning. We asked participants about the ramifications of ignoring the malware warning (e.g., "Which outcome is the most likely if you clicked through the red page to proceed to youtube.com?"), and the answers differ based on the type of destination. Table 4 shows the percentage of participants who think a bad outcome (i.e., "My computer would be infected by malware.") is most likely to occur. Fewer participants believe there will be a bad outcome when the destination is high-reputation ($\chi^2=265.35$, df=1, $p<0.0001$).

6.2 Referrer Reputation

We asked participants to respond to warnings on sites linked from high- and low reputation referrers. We find that the referrer's reputation had only a weak effect on how participants reacted to the warning scenarios. As Table 3 shows, slightly more participants claim that they would ignore a warning on a site linked from a high-reputation re-

Table 4: Perceived risk of ignoring a malware warning

Experiment	Scenario	% Bad Outcome	N
Destination experiment	High-reputation (YouTube)	34%	823
Destination experiment	Low-reputation (blog)	77%	792
Referrer experiment	High-reputation (Facebook friend)	51%	454
Referrer experiment	Low-reputation (lyrics site)	58%	462

ferrer. However, the difference between the two SRCTRs is not statistically significant (p=0.36).

In the open-ended question responses, some participants said that their trust in friends or mistrust of lyrics sites would influence their decision. For example,

> *Malware is dangerous, and most of those lyrics sites are shady*

> *I find it harder to believe [the warning] when my facebook friend just posted it and had no problems.*

> *I presume that visiting youtube from a facebook link would be safe.*

One participant summarized the difference between the Facebook status update from a friend and the lyrics website:

> *This [lyrics] website is less reliable than my friend who posted the link so I don't know if I should trust it.*

Some participants' responses indicated that they were considering both the reputation of the referrer and the reputation of the destination (YouTube). For example:

> *[I] trust youtube, but I don't necessarily trust the lyrics website*

There was also a weak relationship between the reputation of the referrer and participants' risk perception. We asked participants about the ramifications of ignoring the malware warning (e.g., "Which outcome is the most likely if you clicked through the red page to proceed to youtube.com?"), and the answers differ slightly based on the type of referrer. Table 4 shows the percentage of participants who think a bad outcome ("My computer would be infected by malware.") is most likely to occur if they ignore the warning. Fewer people believe there will be a bad outcome when the referrer is high-reputation (χ^2=4.13, df=1, p<0.05). Although the difference is statistically significant, the practical difference between the conditions is small.

Overall, we found little difference between the high- and low-reputation referrer conditions.

6.3 Getting More Information

There are two ways to get more information about a Chrome malware warning. First, the warning includes a "Learn more" link in the last paragraph. This leads to Google's general online security guide (Figure 6). Second, clicking on the "Advanced" link triggers the appearance of a link named "Details about problems on this website." That link leads to a diagnostic page with technical information (Figure 7).

The interactive variant of the destination experiment allowed participants to access additional information about

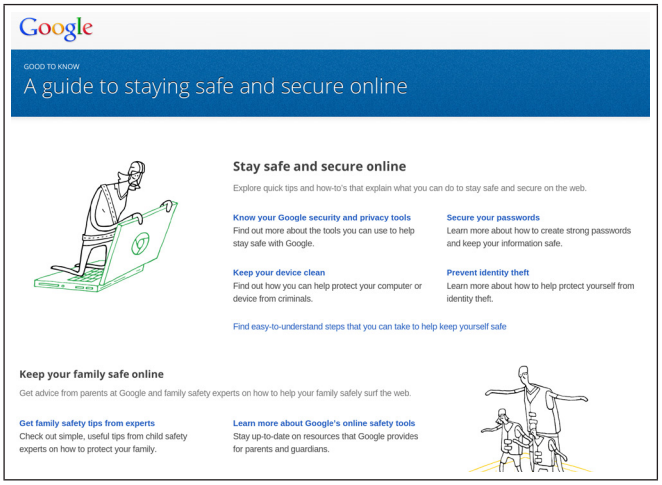


Figure 6: Google's online security guide

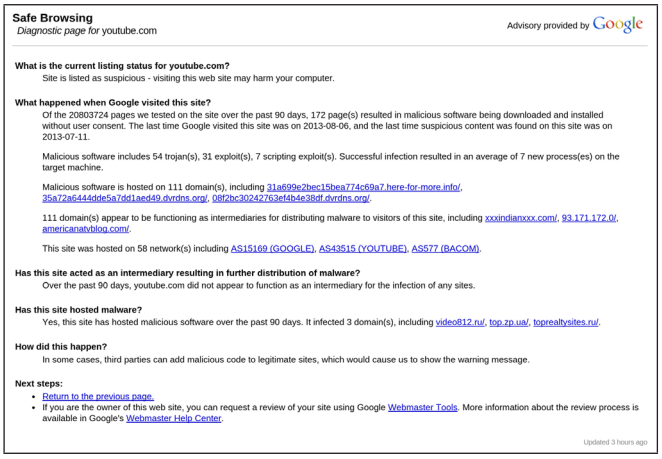


Figure 7: Safe Browsing diagnostic page

warnings before making a decision. Participants were more likely to want additional information in the high-reputation destination scenario. 16% of participants in the interactive condition navigated to the online security guide (6%) or the diagnostic page (9%). In contrast, only 3% of participants sought more information from either source when the warning was for a low-reputation destination.

Unfortunately, participants who saw more information proceeded at the same rate as other participants. Furthermore, participants said they were not satisfied with the content or amount of information on the pages that provide more information. Participants felt that the online security guide was too generic or too lengthy to be helpful:

> *Close out [the page], I would want to know more*

specifically why the warning was brought up for the particular site.

I would probably ignore it and just go to youtube site. There's too much general information on this page for it to be helpful.

It's too much to read.

Although the diagnostic page provides more detailed information, participants were still frustrated by it. Several said that they would look elsewhere. For example,

I would likely not continue, instead I would go to a search engine there and search for the site. This warning is inconsistent with what I believe the integrity of the site is. But, it is possible that this is some sort of advanced hijacking technique.

I would close the tab and check the URL in firefox to see what info I got there. I'd probably also post to twitter and ask if anyone else was getting this info and if so had anyone seen any articles/posts about what kind of malware and who had infected it.

Additionally, several participants in the interactive variant and other variants of the destination experiment indicated that they would seek external information about the warning before making a decision. In particular, participants in the high-reputation destination scenario said they would seek external information from sources such as search engines, news articles, and social media websites.

Something is screwed up, given that it's YouTube. I would search the internet for others reporting the problem.

I would be worried that someone compromised Youtube. I'd try to research and see if this was widespread news (as it likely would be if it were true), or just a problem with Chrome.

I would reenter my search to make sure I didn't click on a link that was masked. If it still showed malware I'd watch news sites to make sure youtube wasn't compromised.

Search the net and find any information on why chrome is blocking youtube.

6.4 Misconceptions

Participants mentioned three notable misconceptions that could hinder the effectiveness of the malware warning.

Protective technology. Some participants believe that they are safe from malware because of protective technology such as anti-virus software or their operating system. E.g.:

I use Linux I'm not afraid of anything.

Because i own a mac and i dont worry about that stuff

I would still proceed knowing I have an anti virus

Other participants had similar responses. These beliefs are dangerous; anti-virus software does not prevent drive-by download attacks, and some drive-by download attacks can succeed on Linux and Mac computers.

Confusion with other warnings. Participants also confused malware warnings with the SSL warning. From their responses, it sounded like they had encountered SSL warnings that they considered to be false positives. For example, one person said:

I know and trust youtube, sometimes the internet browser doesn't have the right certificate.

I want to learn why chrome thinks the site contains malware. Sometimes it might just involve something like an expired security certificate

We also asked participants about prior warnings. About 77% of participants remembered seeing a similar warning in the past. We asked participants to elaborate, and some responses referred to the SSL warning as if it were the same warning. For example:

I believe I got [the warning] because of some discrepancy between http and https.

Identity of the destination site. In the referrer experiment, some participants suspected that the lyrics site might have linked to a site that was not actually YouTube. E.g.:

I don't trust lyrics sites very much, especially ones with those kinds of ads. They could have possibly altered that link to lead to somewhere malicious.

I don't trust redirects from lyric sites.

However, the screenshot in the survey showed a warning for `youtube.com`. The screenshot included the omnibox (which said "`http://www.youtube.com`"), and the malware warning itself includes the destination URL in the text. These participants either did not know how to check the identity of the destination site, did not think to check those identity indicators, or did not trust those identity indicators.

7. IMPLICATIONS

We discuss the implications of our findings and make suggestions for improvement to the warnings. Some of the suggestions have already been adopted by Google Chrome. We also highlight additional open questions and challenges.

7.1 Gaining Users' Trust

Our findings suggest that end users may trust other parties more than they trust the browser's malware warning. In particular, some study participants trusted the reputation of the destination site more than the warning. Some participants also trusted their anti-virus software or operating system to protect them. We recommend adjustments that could increase users' belief in the warnings.

High-reputation destinations. Many participants could not believe that a site like YouTube would be malicious, causing the SRCTR for the high-reputation destination to

be much higher than the SRCTR for the low-reputation destination.[2] Participants' open-ended responses show that this is due to trust in the brand, prior positive experiences with the site, or some combination of the two. Our field data demonstrates that this same effect happens in naturalistic settings for websites that users have previously visited.

We recommend using a special warning for high-reputation destinations. The warning would need to acknowledge that the website is usually safe but emphasize that it has been temporarily compromised. This should match users' mental model better than telling them that the website itself is malicious. One challenge is how to identify high-reputation destinations; a possible solution is to treat all sites in a user's history as high-reputation, combined with a pre-loaded list of high-reputation destinations. Prior literature on site credibility may help guide the identification of high-reputation destinations [8, 16, 18, 22, 27, 39].

How to communicate this information effectively is an open question. The warning already attempts to address this with its third sentence: "Even if you have visited this website safely in the past, visiting it now is very likely to infect your computer with malware." It is not clear whether participants missed this information because they did not read it, or whether they simply found it unconvincing. In our future work, we will be experimenting with different approaches to address this issue.

More information. Our findings suggest that some people are conflicted when they encounter warnings on high-reputation sites and want more information to resolve this conflict. In our study, a notable minority of participants expressed a desire for more information about the warning on a high-reputation destination. There are also ample examples of users asking for more information about malware warnings on web forums and Twitter.

We have already updated the "Learn More" link and diagnostic page in response to this concern. Our participants complained that the general online security guide was too vague, so we modified the "Learn More" link to point to the Safe Browsing Transparency Report. The Transparency Report provides more specific information about why Chrome blocks websites. This change will take effect in Chrome 37. Although the diagnostic page was intended to be more specific, participants found it confusing and unsatisfying. We have built a new version of the diagnostic page that should better address participants' needs. It will be launched in July 2014. Future work is needed to determine whether the new "Learn More" link and diagnostic page will sway undecided users.

Protective technology. Some participants thought that they did not need to heed the malware warning because their anti-virus software or operating system would protect them. Such (often inaccurate) beliefs could expose people to very real risks. We recommend that the warning should specify that neither is an adequate defense against a malicious site.

In the hope of reaching Mac users, Chrome for Mac OS X changes the phrase "your computer" to "your Mac" in the warning. A limitation of our study is that we showed the default PC version ("your computer") to all participants. However, we recommend that this should be made more explicit. People may not notice the subtle reference to Macs.

Role of the referrer. The reputation of the referrer played only a minor role in participants' decisions. We consider three possible reasons: (1) participants do not use the referrer's reputation to make a decision, (2) our experiment lacked the necessary statistical power to identify a small effect, or (3) participants did not consider Facebook statuses to be high-reputation because of the prevalence of Facebook spam. With respect to the third explanation, participants had inconsistent views of the Facebook status:

> *There are always issues like this on facebook. I would not proceed.*
>
> *Someone could have hacked that person's facebook account and posted a false link to a virus.*
>
> *I would trust my friend not to post a bad link but I would be afraid to continue on based on the screen that showed up.*
>
> *from facebook i am less likely to think there is malware associated with the link, especially a youtube link.*

It is possible that more of a difference would appear if the high-reputation referrer were a news website, text message, or other mode of delivery.

We do not have any recommendations to offer about the referrer at this time. However, future work could further investigate the role of different referrers.

7.2 Differentiate Malware and SSL Warnings

Some participants confused Chrome's malware and SSL warnings. This is undesirable because SSL warnings are often false positives; we worry that this devalues user perception of the malware warning. Furthermore, malware warnings put the security and privacy of the whole computer at risk, not just the confidentiality and integrity of a single domain. Ideally, malware warnings should be taken more seriously than SSL warnings.

A possible solution is to make the two warnings more distinct. At the time of our study, both warnings had predominantly red color schemes. We modified Google Chrome's SSL warning to have a yellow-orange background, starting in Chrome 32. In future work, we will investigate if further changes may still be needed to help end users distinguish between the two types of warnings.

7.3 Survey Wording

The type of role and priming with risk information made little difference in participants' responses. Our finding on priming with risk information reinforces similar findings in prior studies [12, 32]. However, interactivity changed participants' responses to our scenarios.

In all but one variant, we asked participants to choose between proceeding and returning to the previous page. In the interactive variant, participants were able to view additional information before deciding. The additional choices significantly decreased the SRCTR in the high-reputation destination condition. However, this was not due to the additional information itself; people who viewed the additional information chose to proceed at the same rate as

[2] We cannot be certain of effect size because the interactive and non-interactive survey variants yielded different gaps between the high- and low-reputation destinations. However, the gap was large in all variants.

other participants. Instead, this suggests that the presence of more choices changed how participants responded to the question. Since we do not know the ground truth for these participants, we do not know whether the interactive or non-interactive variant better represents the participants' real world behavior. Future work could further investigate this effect.

7.4 Open Question: Daily Variance

The malware warning CTR in Chrome fluctuates over time in the field. Discovering the cause of this fluctuation could help warning designers reduce the CTR. Ideally, the warning would be modified to address the situations that lead to sudden increases in CTR.

Prior experience. As discussed in Section 3, we originally hypothesized that the daily variance was due to the daily rate at which familiar websites appeared on the Safe Browsing list. However, our data did not support this hypothesis. Nonetheless, we did discover one clue: the daily variance is larger for warnings on previously visited websites than for warnings on new websites. The daily variance might therefore be related to prior experience with the website. For example, it could be due to the quality of the website or how much the user likes the website.

News stories and social media. Another possible explanation for the daily variance is that high-profile news stories or social media discussions influence users' reactions to warnings. Warnings on popular websites are sometimes mentioned in the news, and we have seen people turn to social media (Twitter, message boards, etc.) to ask each other about warnings on high-profile websites. This might be more likely for previously visited websites, since users might find those warnings more puzzling. Several participants in the Mechanical Turk study said that they would search for more information if they saw a warning for YouTube.

For example, Section 3.1 describes an event on February 9 that was covered in the press and discussed by many on social media. A similar event took place the week before, on February 4, 2013. An advertising network was put on the Safe Browsing malware list because its homepage was compromised. It was initially unclear whether its advertisement serving infrastructure was compromised as well. This caused malware warnings to appear on several high-reputation sites that use the advertising network (e.g., Huffington Post, Washington Post, The New York Times). This event caused the number of warning impressions to dramatically increase: from approximately 100,000 on a "typical" day to 1,254,520 on February 4 (within the subset of the population that shares statistics with Google).

The two events on February 4 and February 9 were fairly similar. Both led to malware warnings on popular websites, made the news, and swamped social media websites. However, users responded differently to the two events: they clicked through only 8% of warnings on February 9 but 15% of warnings on February 4. What was different? On February 9, news stories and social media posts exhorted users to heed the warning. Users saw the recommendation, and the CTR decreased to 8%. In contrast, the advertising company involved in the February 4 event issued a statement saying that the warning was a "false alarm" [28], and news outlets reported that the warnings were false positives [20, 29].

Anectodal evidence is insufficient to substantiate a hypothesis, but the role of news stories and social media should be investigated further. Measuring the influence of news and social media on user behavior is left for future work.

8. CONCLUDING SUMMARY

Our goal is to understand why users ignore malware warnings. To this end, we analyzed 3,875,758 Chrome malware warning impressions from the field and ran an online, survey-based controlled experiment with 1,397 participants.

We found that users in the field are twice as likely to ignore a malware warning from Chrome if the blocked website is already in their browsing history. This suggests that users are less likely to believe a malware warning if they have prior experiences with a website. Participants in our online study echoed this sentiment: they said that they did not believe that a popular, high-quality site could be malicious. Furthermore, participants' quotes indicated that some people have misconceptions about the warning; for example, some participants confused the malware and SSL warnings.

Our primary recommendation is that malware warnings need to be changed to convey that high-reputation websites can be temporarily compromised. This will address the unfortunately common situation where malware authors take control of popular websites to spread malware. Some participants also expressed a desire for clear, contextual information to help them make a decision. To address this latter concern, we adjusted the Chrome warning's "Learn More" link and built a new diagnostic page. Our work on improving the Chrome malware warning continues.

Data Collection Ethics

All Chrome metrics are subject to privacy policies. Participants opt in to share statistics with Google, and participants can later opt out by changing their settings [19]. Our new statistics were reviewed according to the Chromium review process. We did not collect or analyze any private or personally identifiable information.

Our Amazon Mechanical Turk experiment asked participants about hypothetical scenarios and prior warning encounters. None of this data is private or sensitive. We also collected optional demographic information on age, gender, social media usage, and education. Our institution does not have an Institutional Review Board (IRB), so it was not subject to IRB review; however, multiple researchers who have received human subjects training reviewed the survey instrument prior to the experiment. We paid the study participants (Amazon Mechanical Turk workers) a rate intended to mimic California's minimum wage.

9. REFERENCES

[1] Desitvforum.com Site Info. http://www.alexa.com/siteinfo/desitvforum.net. Accessed: 2013-11-7.

[2] Php.net Site Info. http://www.alexa.com/siteinfo/php.net. Accessed: 2013-11-7.

[3] Safe Browsing API. https://developers.google.com/safe-browsing/. Accessed: 2013-11-9.

[4] Warriorforum.com Site Info. http://www.alexa.com/siteinfo/warriorforum.com. Accessed: 2013-11-7.

[5] D. Akhawe and A. P. Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the 22th USENIX Security Symposium*, 2013.

[6] F. Asgharpour, D. Liu, and L. J. Camp. Mental models of computer security risks. In *WEIS*, 2007.

[7] P. Briggs, B. Burford, A. De Angeli, and P. Lynch. Trust in online advice. *Social Science Computer Review*, 20(3):321–332, 2002.

[8] P. Briggs, B. Burford, A. De Angeli, and P. Lynch. Trust in online advice. *Social Science Computer Review*, 20(3):321–332, 2002.

[9] R. Dhamija, J. Tygar, and M. Hearst. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[10] S. Egelman, L. F. Cranor, and J. Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1065–1074, New York, NY, USA, 2008. ACM.

[11] S. Egelman and S. Schechter. The importance of being earnest [in security warnings]. In *The International Conference on Financial Cryptography and Data Security*, FC '13, 2013.

[12] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On The Ecological Validity of a Password Study. In *Proceedings of the Symposium On Usable Privacy and Security*, 2013.

[13] A. P. Felt, R. W. Reeder, H. Almuhimedi, and S. Consolvo. Experimenting at scale with google chrome's ssl warning. 2014.

[14] R. J. Fisher. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*, 20(2), September 1993.

[15] B. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, et al. What makes web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68. ACM, 2001.

[16] B. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, et al. What makes web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68. ACM, 2001.

[17] B. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber. How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15. ACM, 2003.

[18] B. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber. How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15. ACM, 2003.

[19] Google Chrome support. Usage statistics and crash reports. https://support.google.com/chrome/answer/96817?hl=en.

[20] R. Greenfield. Why Malware Warnings Took Over the Internet Today. http://www.theatlanticwire.com/technology/2013/02/google-chrome-malware-warnings/61774/, February 2013.

[21] J. Haber. SmartScreen Application Reputation in IE9, May 2011.

[22] J. Kim and J. Y. Moon. Designing towards emotional usability in customer interfaces–trustworthiness of cyber-banking system interfaces. *Interacting with computers*, 10(1):1–29, 1998.

[23] J. Kim and J. Y. Moon. Designing towards emotional usability in customer interfacesÑtrustworthiness of cyber-banking system interfaces. *Interacting with computers*, 10(1):1–29, 1998.

[24] T. H.-J. Kim, P. Gupta, J. Han, E. Owusu, J. Hong, A. Perrig, and D. Gao. Oto: online trust oracle for user-centric trust establishment. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 391–403. ACM, 2012.

[25] E. Limer. Google Chrome Is Blocking a Bunch of Major Sites for Malware, Even YouTube, February 2013.

[26] McAfee. The web's most dangerous search terms. Report, 2009.

[27] M. J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

[28] NetSeer. Twitter post. https://twitter.com/NetSeer/status/298498027369402368, February 2013.

[29] J. C. Owens. Google Chrome's NetSeer malware warning blocks websites, company says no virus distributed. http://www.mercurynews.com/ci_22515730/malware-warning-citing-netseer-blocks-google-chrome-users, February 2013.

[30] N. Provos. Safe Browsing - Protecting Web users for 5 Years and Counting. http://googleonlinesecurity.blogspot.com/2012/06/safe-browsing-protecting-web-users-for.html, June 2012.

[31] N. Provos, D. McNamee, P. Mavrommatis, K. Wang, N. Modadugu, et al. The ghost in the browser analysis of web-based malware. In *Proceedings of the First Workshop on Hot Topics in Understanding Botnets*, 2007.

[32] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The Emperor's New Security Indicators. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 2007.

[33] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010.

[34] K. Šolic and V. Ilakovac. Security perception ot a portable pc user (the difference between medical doctors and engineers): A pilot study. *Medicinski Glasnik*, 6(2), 2009.

[35] A. Sotirakopoulos, K. Hawkey, and K. Beznosov. On the Challenges in Usable Security Lab Studies: Lessons Learned from Replicationg a Study on SSL Warnings. In *Proceedings of the Symposium on Usable Privacy and Security*, 2011.

[36] T. C. Sottek. Malware warnings ripple across the web just five days after last major incident. `http://www.theverge.com/2013/2/9/3971766/major-websites-hit-with-malware-warning`, February 2013.

[37] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *Proceedings of the USENIX Security Symposium*, 2009.

[38] Y. Wang, D. Beck, X. Jiang, R. Roussev, C. Verbowski, S. Chen, and S. T. King. Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites That Exploit Browser Vulnerabilities. In *Proceedings of the Network and Distributed System Security Symposium*, 2006.

[39] C. N. Wathen and J. Burkell. Believe it or not: Factors influencing credibility on the web. *Journal of the American society for information science and technology*, 53(2):134–144, 2002.

[40] T. Whalen and K. M. Inkpen. Gathering Evidence: User of Visual Security Cues in Web Browsers. In *Proceedings of the Conference on Graphics Interfaces*, 2005.

[41] M. Wu, R. C. Miller, and S. L. Garfinkel. Do Security Toolbars Actually Prevent Phishing Attacks? In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2006.

APPENDIX

The following is a full example of the destination experiment survey, using the "personal" wording variant.

Mechanical Turk HIT Description

How do you browse the web?

We are conducting a survey about how you browse the web. This survey will ask you about how you would react to different situations on the web. The whole survey should not take more than 10 minutes. Please answer each question carefully and answer honestly. We will pay you $1 for your participation.

To participate:

1. You must be 18 years old or older.
2. You must be a Chrome user.
3. You must be in the United States while taking the survey.
4. You must be an English language speaker.
5. You must NOT participate in the survey more than once.

To be paid, follow these steps:

1. Click on the link below to go to the survey:
2. The link will appear here when you accept this HIT.
3. After completing the survey you will receive a confirmation code in the last page.
4. Enter the code in the box below and we will approve your payment.
5. Please enable Javascript to perform this HIT.

Enter code here: [ ]

For questions and problems, please contact us through Mechanical Turk's contact functionality.

Thank you!

Researchers at Google

Page 1

How do you browse the web?
Thank you for your interest in participating in our survey. Please click "Continue" to start the survey.

Page 2

Which of these websites have you heard of? (check all that apply)

- Diaspora
- Facebook
- FunFactsOfLife
- SnackWorks
- Vimeo
- Wikipedia
- YouTube

Page 3

Imagine that you are visiting www.facebook.com to check friends' latest updates. You see the status update shown below.
[Figure 4(b)]
Which band recorded the song shown in the status update?

- Four Men
- Weezer
- The Beatles
- The Clash

After clicking on the link to watch related videos, you see the page shown below.
[Figure 1]
What would you do?

- Proceed to youtube.com.
- Go back (do not proceed to youtube.com).
- Other. Please specify:

Why? (short-essay)

Page 4

Imagine that you are visiting www.facebook.com to check friends' latest updates. You see the status update shown below.
[Figure 4(c)]
What is the name of the blog shown in the status update?

- Monkeys
- TechCrunch
- The Fast Runner
- Fun Facts Of Life

After clicking on the link to read the full blog post, you see the page shown below.
[Figure 1, but with the blog as the URL]
What would you do?

- Proceed to [blog URL]
- Go back (do not proceed to [blog URL]).
- Other. Please specify:

Why? (short-essay)

Page 5

Which outcome is the most likely if you clicked through the red page to proceed to youtube.com?

- I would be able to watch videos with no malware.
- My computer would be infected by malware.
- Other. Please specify:

Which outcome is the most likely if you clicked through the red page to proceed to [blog URL]?

- I would be able to read the blog post with no malware.
- My computer would be infected by malware.
- Other. Please specify:

Page 6

How would you typically react if you saw a similar red page when trying to visit a website in your day-to-day life?

- I would typically proceed to the website.
- I would typically go back (wouldn't proceed to the website).
- Other. Please specify:

Page 7

Before this survey, had you ever seen a similar red page when trying to visit any website?

- Yes
- No
- I don't remember

If the respondent chooses "Yes", then:
What happened the last time you saw a similar red page when trying to visit a website? (What was the website? What did you do?) (short essay)

Page 8

In the past week, how many times have you checked Facebook?

- I have never heard of Facebook.
- I have heard of Facebook but I do not have a Facebook account.
- Zero times in the past week
- Once in the past week
- Twice in the past week
- Three times or more in the past week

In the past week, how many videos have you watched on YouTube?

- I have never heard of YouTube.
- None in the past week
- 1 video in the past week
- 2 videos in the past week
- 3 or more videos in the past week

What would you do if your wireless router at home were not working?

- I do not know what a wireless router is.
- I would call the provider's technical support to fix it.
- I would call a friend to help me to fix it.
- I would fix it myself.
- Other. Please specify:

What is a computer firewall?

- I do not know what a computer firewall is.
- Software that locates the nearest fire station.
- Software that encrypts personal files.
- Software that controls network traffic to/from a computer.
- Other. Please specify:

What is a public key certificate?

- I do not know what a public key certificate is.
- An electronic document that shows a computer is virus-free.
- An electronic document that shows a website is using 2-factor authentication.
- An electronic document that shows the identity of a website.
- Other. Please specify:

Page 9

What is your gender?

- Male
- Female

What is your age? (free response)
What is your highest completed level of education?

- Professional doctorate (e.g., MD, JD, DDS, DVM, LLB)
- Doctoral degree (e.g., PhD, EdD)
- Masters degree (e.g., MS, MBA, MEng, MA, MEd, MSW)
- Bachelors degree (e.g., BS, BA)
- Associates degree (e.g., AS, AA)
- Some college, no degree
- Technical/Trade school
- Regular high school diploma
- GED or alternative credential
- Some High School
- Other. Please specify:

Which operating systems do you normally use? (check all that apply)

- Windows
- Mac OS
- Linux
- iOS
- Android
- I don't know
- Other. Please specify:

Which web browsers do you normally use? (check all that apply)

- Microsoft Internet Explorer (IE)
- Mozilla Firefox
- Google Chrome
- Apple Safari
- Opera
- I don't know
- Other. Please specify:

Which web browser do you use the most on your personal computer(s)?

- Microsoft Internet Explorer (IE)
- Mozilla Firefox
- Google Chrome
- Apple Safari
- Opera
- I don't know
- Other. Please specify:

Page 10

If you have any additional comments, please write them here. (short essay)

Page 11

Please copy the following code and paste into the text box in the HIT before clicking "Submit".
Check that this is your Amazon worker ID
Submit

Exploring Internet Security Perceptions and Practices in Urban Ghana

Jay Chen
NYU Abu Dhabi
PO Box 129188
Abu Dhabi, UAE
jchen@cs.nyu.edu

Michael Paik
NYU Abu Dhabi
PO Box 129188
Abu Dhabi, UAE
mpaik@cs.nyu.edu

Kelly McCabe
NYU Abu Dhabi
PO Box 129188
Abu Dhabi, UAE
kellymccabe@nyu.edu

ABSTRACT

Security is predicated, in part, upon the clear understanding of threats and the use of strategies to mitigate these threats. Internet landscapes and the use of the Internet in developing countries are vastly different compared to those in rich countries where technology is more pervasive. In this work, we explore the use of Internet technology throughout urban and peri-urban Ghana and examine attitudes toward security to gauge the extent to which this new population of technology users may be vulnerable to attacks. We find that, like in North America and Europe, the prevalent mental threat model indicates a lack of understanding of how Internet technologies operate. As a result, people rely heavily upon passwords for security online and those who augment their security do so with a variety of ad hoc practices learned by word of mouth. We relate and contrast our findings to previous works and make several recommendations for improving security in these contexts.

Keywords

ICTD; Security; Passwords; Facebook; Google; WhatsApp; Social Networks; Ghana

Categories and Subject Descriptors

H.5.m. [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

1. INTRODUCTION

Users in the developing world face a significantly different Internet landscape than users in rich countries. Connectivity can be poor or absent, understanding of how technologies work can be *ad hoc* without any systematization due to lack of exposure, and threat models can be both different and poorly understood. Relative to rich countries, developing countries have may have substantially less training and experience with Internet technologies [18]. Internet penetration and therefore use are on the rise in the developing world, and in Ghana in particular [13] and it is possible that the uptake of Internet technologies will soon outgrow commonly held security attitudes and commonly practiced security measures.

Networked security has historically been an "arms race" between intruders becoming more sophisticated and security experts rushing to defend against the latest exploits. The battleground has thusfar mostly been isolated to rich countries and large corporations, but as the GDP of countries like Ghana increases [5], these countries become more attractive targets. Furthermore, because the threats can be very advanced compared to the local experience in developing countries, these populations may be especially vulnerable to attacks. This scenario is especially worrying because for many developing countries networked infrastructures are being increasingly relied upon for critical services such as mobile banking, e-health, and e-government [45, 54].

In order to prevent such worst-case scenarios, we need to develop better technologies and improve awareness. Before this, we should understand people's existing perceptions of technology, people's mental models of networked security, and how they defend against threats. To understand the current security environment, we conducted a study to understand the specific use cases and the rationale that people in Ghana rely on to make decisions about their security practices. We conducted surveys and interviews of 193 respondents across 8 regions in Ghana focused on capturing users' perceptions, practices, and experiences. Our contribution is to provide information about the use of the Internet by urban Ghanaians and their perceptions of and measures for maintaining Internet security.

Wash [53] recently studied mental models of home computer security in an attempt to understand how home users make security decisions. Here, our emphasis is not to build distinct categorizations, but instead to gather salient features from asking two basic research questions: 1) Perception of threats: How do Ghanaian Internet users perceive security threats online and how confident are they in their ability to protect themselves? 2) Security measures: What measures do Ghanaian Internet users employ to protect themselves from online threats?

We find that confidence with Internet technologies is relatively high, particularly for mobile phones. Unfortunately, we also find that certain security behaviors are quite lax, and are often based on misconceptions or mischaracterizations of how technologies work. In particular we discovered that terminology regarding threats were often conflated and that the use of passwords is generally seen as an all-encompassing panacea. As a result, all manner of private information is held behind a security model based solely upon passwords. We further find that users are typically only concerned with immediate, local, physical threats in the form of people who may come to the terminal that they had been using and try to extract information from it; threats from the network side, whether from malicious sites posing as innocuous ones or between users and the

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

sites they are using, were not part of users' mental threat model. While these results are troubling, the low incidence of local experience with hacking suggests that this mental model and corresponding security measures taken may be entirely rational.

In the remainder of the paper, we first discuss related work in technology use in developing countries, conventional security perceptions and models as observed in the U.S., and the relevant security countermeasures. We then detail the methodology of our study and our findings. From these findings, we propose several ideas for potential mitigations, suggest ways to educate users, and enumerate avenues for future research.

2. BACKGROUND AND RELATED WORK

To lay the groundwork for our research we discuss some related work and motivating reasons for studying networked security in a developing country like Ghana.

2.1 Understanding Internet Security

Managing computer security is a challenging task and has been studied extensively in the past in conventional contexts such as the home, workplace, and public areas [23, 30, 36, 53]. Dourish's work exploring user attitudes toward computer security in developed countries have revealed that people generally perceive security as frustrating barriers to productivity and ultimately futile [23]. Dourish and Grinter found that users typically delegate security to the technology itself, other individuals, entities, or organizations [23, 30]. Herley argues that users' rejection of the security advice they receive is entirely rational from an economic perspective [31].

Research from e.g. Lindgaard *et al.* [39] and Cyr *et al.* [22] clearly demonstrates that the trustworthiness of a website is dependent, at least in some ways and to some degree, on the way it is presented to the user and the user's perception of its quality. People have been designing webpages with this in mind for at least 15 years, (e.g. Kim and Moon [35]). Research by Everard *et al.* [26] also shows that site presentation flaws can also affect trustworthiness. This phenomenon has also been studied and modeled across cultures by e.g. Cyr *et al.* [21, 22], though cultural impact is less well understood in the developing world. Jakobsson *et al.* find that trustworthiness often relies on cues *not* designed as security features [33].

The perception of threats is a complex problem, as shown by a survey of this research space. Psychological research (e.g. [28]) illuminates this question somewhat, showing that people learn about threats if the perception of the threat is perceptually correlated to confirmatory information, but it is less clear how physical disconnectedness and mental world models correlate with this perception. Recent work by Wash and Rader show the mental models non-expert computer users rely on to make security decisions [46, 53]. They find that much of the knowledge of non-expert computer users is gleaned from stories that act as informal lessons about security. In developing countries because anti-virus software is relatively expensive and formal computer training is less available these perceptions and behaviors may be more dependent upon these informally learned strategies.

2.2 Internet Landscape in Ghana

Prior work in Ghana by Burrell focuses for the most part on computer use in Internet cafes [19, 20]. Burrell found widespread use of Internet (in Accra) and prevalence of social networking and chat services (as well as voice calls) to reach out to foreign and domestic contacts [27]. Online social network use and chatting are widespread across Africa and developing countries elsewhere. Research by Wyche *et al.* [54] has extensively explored the use of social networks in Nairobi, Kenya. Wyche finds that the users she studies in a Nairobi slum use Facebook in myriad relatively sophisticated ways including the creation of fan pages to promote businesses, sharing film, photos, and audio, actively soliciting friends for work, and the like [54, 55]. In our user group, casual chat with friends was the primary and many times only use people had for Facebook, and the uptake of WhatsApp (which essentially provides only social chat functionality) is consistent with this.

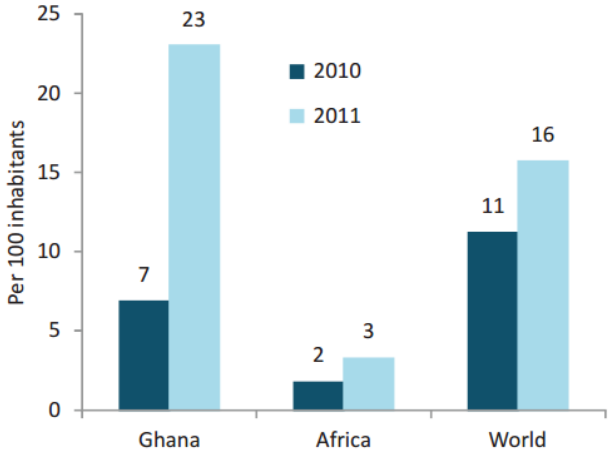


Figure 1: Active mobile-broadband subscriptions per 100 inhabitants, 2010-2011, Ghana in comparison with regional and world average. From [13].

We found that the mobile Internet penetration rate was significantly higher than numbers reported in 2012, in line with strong year-over-year growth. Figure 1 shows data from the International Telecommunications Union indicating that Ghana saw approximately 23% mobile broadband penetration in 2011; in our sample from Ghana the penetration rate was well over 50%, with iPhones, Android and Windows Mobile phones, and data-enabled feature phones all represented. All of the respondents were encountered in urban or peri-urban environments, so this number likely trends high since mobile data coverage and, therefore, penetration drops off precipitously in rural areas. This does, however, illustrate the dramatic progress in mobile data uptake in Ghanaian urban areas and the impending need for usable security tools.

2.3 Internet Security in Developing Countries

Specific to security in developing countries, Ben-David *et al.* have found that technology users face a complex set of security concerns that are deeply tied to a range of contextual factors that make importing security solutions from industrialized countries inadequate [16]. The specific factors that make the problem especially challenging in developing regions include: poor security hygiene due to scarce bandwidth and frequent network failures [44, 54], unique usage patterns (e.g. reliance on non-standardized protocols for mobile banking [45], and shared use of PCs [18]), software piracy [15, 34], and novice users [47]. In terms of security solutions, however, only a few security mechanisms have been designed for developing region contexts [43, 45].

2.4 Passwords and Other Security Mechanisms

Passwords and studies of passwords have been around nearly as long as computer accounts have, as shown by Morris and Thomp-

son's 1979 paper [41]. Passwords can be weak due to human factors, but there is no clear evidence about how stronger passwords actually help [29]. It has been well demonstrated that people do not like to change their passwords very often [32], despite the potential risk passwords weakening over time in the face of increased attacker sophistication and of accidental password exposure. Password strength meters are generally ineffective as people ignore them and changing this behavior is difficult [25, 31, 52]; some sites (e.g. Microsoft accounts) have found it very effective to simply ban popular passwords [7]. In a study by Kuo *et al.* in which users were instructed to use mnemonics, the great majority of passwords in the study generated using mnemonics could actually not be guessed [38]. It is unclear whether users who have had less exposure to hacking choose passwords that are less resistant to attack and whether they should be inculcated with password 'best practices'.

Of additional concern for our userbase, users have been shown by Sun *et al.* to be unable to distinguish real and fake Google login forms even when prompted [50], making the use of passwords potentially less secure. Forcing users to follow best practices is an option, and generally people find it irksome, but feel safer [48]. However, forcing onerous security upon users has been demonstrated to cause them to find ways to circumvent that security [42]. Furthermore, in contexts where hacking is rare, it is especially unclear whether following additional security precautions is actually a rational decision when even in developed countries following best practices may not be a rational decision [31]. Two-factor authentication as recently implemented by Facebook, Google, and Yahoo [3, 6, 12] may be less useful for Ghanaians because a large proportion of users in Ghana are using these services only from their mobile phone. Many other mechanisms such as notifications and browser popups have been proposed by mainstream security and privacy researchers, but user habituation can erode the effectiveness of such methods [24, 37].

3. SETTINGS AND METHODOLOGY

We conducted a qualitative study of how technology users use the Internet and think about security. We used surveys and semi-structured interviews to conduct our research. We conducted 193 surveys and interviews during Summer 2013 and we conducted our analysis in October 2013. Nearly all respondents were surveyed on Fridays, Saturdays, and Sundays. All surveys and interviews were conducted in English (the official language of Ghana) and the interviews were digitally recorded. Interviews averaged 10 minutes each and they were audio recorded and transcribed for analysis. Standard procedures for informed, voluntary consent were practiced. Users were offered a 10 Ghanaian Cedi payment (approximately 5 USD) to participate in the study, and were instructed that they could discontinue taking the survey or refuse to be interviewed at any point (13 respondents opted not to be interviewed and 7 did not wish to complete the demographic information), and still receive this payment. Interviews were conducted by one Ghanaian male and one white American female. Our analysis does not indicate any bias in content of responses correlating to the race or gender of the interviewer.

Respondents were chosen from a sample of technology users we encountered in public gathering places such as streets, Internet cafés, markets, and universities on weekends to select for maximum variation. The respondents were gathered from across 11 urban and peri-urban locations in 8 geographical regions in Ghana. Table 1 lists the locations and settings where we gathered respondents. We began by screening these potential respondents to exclude people who had no experience with mobile phones or computers and those below the age of 18. Ages ranged from students (18 years old) up through executives (55 years old), but the vast majority were between 18 and 31 (the median age was 25).[1] There were 131 male (68%) and 55 female (28%) respondents.

Region/City	Location	# Resp.
Accra-Osu	street and copy center	11
Accra-Nima	street and restaurant	10
Accra-Airport	office	4
Accra-Kokomlemle	Internet café	29
Eastern Region-Korforidua	street and a college	20
Northern Region-Tamale	community event	23
Volta Region-Ada	street	15
Central Region-Takoradi	street outside a market	20
Ashanti Region-Kumasi	college and a street	21
Brang Ahafo-Sunyani	streets	19
Western Region-Cape Coast	university	20

Table 1: Number of respondents by region/city and location.

Education Level	# Resp.
Junior secondary school or less	13 (7%)
Senior secondary school	53 (28%)
Polytechnic or post-secondary teacher training	37 (20%)
University	63 (34%)
Graduate school	20 (11%)

Table 2: Number of respondents by highest education level.

From those not excluded, we selected respondents for maximum diversity by choosing respondents from a wide variety of backgrounds, ages, and socio-economic classes. Socio-economic status was not explicitly measured in terms of income, but occupations ranged from service industry workers (cook, hairstylist, etc.) up to professionals (IT professionals, engineers, etc.). Table 2 lists the education level and Table 3 lists the occupations of our respondents. We believe that our sample is fairly representative of the urban and peri-urban population of Ghana and allowed us to document diverse variations in attitudes toward technology and perceptions of security to identify important patterns. Figure 2 illustrates one street-side interview taking place in Kumasi.

We developed a survey and a face-to-face semi-structured interview protocol that explores several aspects of the use and attitudes toward Internet security. Our interview participants were the subset of surveyed respondents who agreed to an interview. In our interviews we specifically probed for instances where respondents encountered hacking or security indications in their interactions. The majority of the interview was spent on asking questions from a pool of questions regarding potentially risky use of technology, awareness of security precautions, and perceptions of security indicators on the Internet. We probed deeper into the responses of the subject when particularly novel responses were given. This method allowed us get a broad picture of the self-reported reasoning behind certain behaviors and attitudes.

The focus of our interviews was exploratory. We asked about incidents or stories regarding hacking and about precautions on the Internet both in terms of security and privacy. We also asked about mobile phone, website, and pen-drive use. We probed deeper into each of these areas to find out the indicators that respondents used to mitigate risk (e.g. appearance of websites, padlock icon on

[1] The median age in Ghana in 2013 was 20.7 years old [1].

Occupation	# Resp.
Student	46 (24%)
Service Industry	15 (8%)
IT / Engineer	12 (6%)
Teacher	10 (5%)
Administrative / Clerical	9 (5%)
Film / Design	5 (3%)
Business / Entrepreneur	5 (3%)
Mobile Banker	3 (2%)
Farming	2 (1%)
No Response	62 (32%)
Other	24 (12%)

Table 3: Number of respondents by occupation.

Figure 2: Respondents filling out surveys in Kumasi.

browsers, etc.).

After collection and transcription of the data, two of the co-authors coded the data independently to look for predetermined and emergent themes. We then discussed these major themes among all of the three co-authors to validate the themes and then expanded themes and organized them into a unified data matrix to identify patterns across subjects and check for representativeness. We used this data matrix to highlight specific examples of trends that appear as descriptions throughout the paper.

4. FINDINGS

We received a total of 193 completed surveys from our respondents and completed 178 interviews. We elaborate on these results below, and believe that together, these results illustrate how people use and perceive technology, what people's attitudes attitudes and perceptions are like with regards to security and privacy, and the measures that people take toward securing themselves.

4.1 Technology Use and Perceptions

All respondents used mobile phones and owned an average of 1.93 sim cards. 184 respondents used the Internet. The survey data from Table 2 and Table 3 show a wide variety of education levels and a levels occupations. Table 4 shows the locations where our respondents accessed the Internet. Our respondents generally accessed the Internet from their personal mobile phones (72%) followed by computers at home (50%), Internet cafés (40%), and

Location	# Resp.
On personal mobile	139 (72%)
Computer at home	97 (50%)
Internet café	78 (40%)
Computer at school	70 (36%)
On other mobile	20 (10%)

Table 4: How the Internet is accessed by location.

Use case	Internet	On Mobile
Facebook/Social Networking	67%	58%↓
Searching	63%	60%↓
Email	59%	65%↑
News	58%	64%↑
Education	58%	53%↓
Entertainment	48%	60%↑
Job Search	22%	21%↓
Games	33%	62%↑
Health	25%	25%
Video/Audio Chat	24%	34%↑
Banking	11%	20%↑
Instant Messaging	9%	14%↑
Agricultural	4%	7%↑

Table 5: How the Internet is used in general and on mobile phones. Arrows indicate increase or decrease on mobile phones compared to general use.

schools (36%). Table 5 summarizes the reported purpose of using the Internet by our respondents. Our findings indicate that the Internet is generally used for social networking, searching, email, news, education, and entertainment. These numbers are quite high, but generally consistent with recent notable findings on the popularity of online social media, job search, and branchless banking in Ghana and elsewhere in sub-Saharan Africa [19, 20, 27, 40, 49, 54]. We were surprised by some of these results, particularly how many people used the Internet for health services (25%).[2]

We asked a number of questions about self-reported skill with computers and mobile phones along with general attitudes and perceptions of security and privacy. Respondents rated their responses on a 5-point Likert scale. On average our respondents reported higher mobile phone skill (4.0) than computer skill (3.1). Of our respondents, 48.7% reported that they more than 5 years of experience using the Internet, 14.5% had 3-5 years of experience, 13.5% 1-3 had years of experience, 10.3% less than 1 year of experience, and 6.3% never used the Internet (6.7% did not respond to this question). Most of the respondents who had never used the Internet had junior secondary school or less levels of education.[3]

4.1.1 Social Networking and Chat

Social networks were extremely popular among our survey group and were clearly a primary reason to go online. From our survey we found that social networking was used by 67% of our respondents and 58% of our respondent on mobile phones. In our interviews, Facebook was mentioned by nearly 30% of respondents

[2]We discuss in detail why some of these numbers are subject to interpretation in Section 4.1.2.

[3]Nine respondents who responded that they never used the Internet responded that they used Internet services. We included those respondents in our results and discuss this issue in Section 4.1.2.

when asked what they do most frequently online (though some were asked specifically whether they had Facebook accounts), with WhatsApp Messenger second most frequently mentioned and Google+, Twitter, Yookos, Yahoo Messenger, and unspecified social networks and messaging applications trailing far behind. Among users of Facebook, person-to-person and group chat were far and away the most mentioned features used; in the words of one respondent:

Interviewer: *What do you do on Facebook?*
Respondent: *I chat.*

and another:

Interviewer: *What do you do on Facebook?*
Respondent: *Facebook? I chat with my friends. And my family.*

Four users of Facebook reported that Facebook was the *only* reason they went online, that they didn't visit any other sites, e.g.:

Respondent: *It's just Facebook, that's all.*

Chatting appeared, for most users of Facebook, to be the only reason to use the site, with users chatting with friends, colleagues, and customers. Of those who chatted with friends, the attraction of chat seemed not to be the ability to convey particular information or to reach particular friends on demand, but to have casual ad hoc chats; that is, it was more important to chat with *someone* than with anyone in particular. Of the dozens of people who had Facebook accounts, during the interviews only two mentioned posting or commenting, one mentioned music and movies, and one mentioned photos; this again despite several users going online only to use Facebook, leading us to believe that many were therefore ultimately going online only to chat. In addition, we found during the interviews that chatting was the most commonly reported use of mobile phones after calling, SMS, and web browsing. This is in contrast to the survey results, which indicate much lower numbers likely because chat was often folded into the responses for social networking and Facebook.

This predilection for chat helps to account for the relative popularity of WhatsApp Messenger [11]. WhatsApp is a cross-platform free messaging app supported by nearly every phone platform and offering free unlimited messaging for the first year. Mentioned, unprompted, by 9 respondents and used by many others, the uptake of WhatsApp, boasting 300 million monthly active users worldwide as of August 2013 [10] (compared to Facebook's 1.15 billion monthly active users [4] as of July 2013), was a surprise. Users of WhatsApp in our survey group reported using it for person-to-person chat as well as for group chat, with at least one user reporting using this group chat feature for work:

Respondent: *I do spend a lot of time, maybe on WhatsApp. Because I'm a media man, and normally we use to discuss, we have a crew page over there we use to discuss concepts we are about to shoot [unintelligible], so normally I'm on WhatsApp.*

Moreover, several users indicated an increasing preference for WhatsApp over Facebook, though the reason for this is not made clear:

Respondent: *Nowadays WhatsApp. So, Facebook has become a little bit, yeah, down, so I do WhatsApp in most.*

If chat is the "killer app" for this user group, it stands to reason that as the cost of data on mobile platforms decreases and its availability increases, the always-on nature of WhatsApp messaging vis-à-vis having to log into Facebook or another social network site on the web will make it increasingly attractive. This is particularly the case as text-based chat, which was the only use of WhatsApp mentioned, consumes paltry amounts of data and is therefore less sensitive to the speed of the underlying data connection, something that is untrue for media-rich sites such as Facebook (though the mobile-friendly version of Facebook improves upon this).

We surmise that the reason that the homogeneity of social networks (nearly all Facebook of those who specified) and chat applications (nearly all WhatsApp) reflects the positive externalities of network effects in tandem with the relative expense and slowness of Internet access: as the number of participants in a network increases, the value of that network increases, and on an expensive connections, users will tend to optimize by only visiting those few sites that provide them the most value per access. For instance, visiting Google+ in addition to Facebook might allow a user to connect with a few more friends, but would incur double the cost in data. Users, therefore, have tended to gravitate towards one or two select sites or applications in each domain of Internet use.

4.1.2 Conflation of Network Services

The interviews revealed several general trends around Internet use and perceptions that we found interesting. One such trend is that among those who mentioned during interviews, unprompted, using the search engine Google, more than 58% specifically referred to it as an educational or research website rather than a general web portal or search engine:

Respondent 1: *Educational websites, go there to research, like Google, yeah?*
Respondent 2: *I go to Google to search for information - I use it to learn.*
Respondent 3: *If it comes to education, I try with Google.*

No respondents indicated that they used Google for any non-educational purpose, such as searching for entertainment, media, or even for news.

Many of the respondents calling Google a research or education site were students at various levels of education, and one specifically mentioned Google Scholar, but the group also included various professionals and at least one person who was unemployed. However, only just over 25% of respondents asked to name the things they do online most frequently mentioned Google.

One result of the spread of mobile connectivity is that for an increasing number, phones are the primary way in which people use network services over alternatives such as Internet cafés. Users expressed reasons such as immediacy and convenience as motivating factors; travel time was also a factor.[4] One user in particular highlights this trend:

Interviewer: *So, do you not usually go to the Internet café?*
Respondent: *No, I don't usually go there.*
Interviewer: *Why is that?*
Respondent: *[chuckles] I don't have the time!*

A direct result of this shift from fixed-line, computer-based Internet use to immediate, mobile, phone-based use is that rather than having a clear delineation between use of the Web and use of other Internet-enabled applications such as instant messaging or phone-

[4] As we will see in Section 4.2.4, perception of insecurity is also a factor.

based apps, users tend to conflate all Internet activities into a class of activities that require data plans on their mobile phones. Thus on the one hand, whereas a user in the United States might say that they are browsing the web *on their phone*, the same idea is expressed without this modifier among our respondents. On the other hand, whereas a user in the US might say that they are using the Facebook *app*, our respondents simply say that they are using Facebook. To our respondents these modifiers appear to be differences without distinctions; the content dictates the label and people appear agnostic to the mode of access, whether from a café or a phone, via an app, or a browser.

4.2 Security

The primary focus of our survey and interviews was to evaluate commonly held security practices and attitudes among respondents, and the interviews revealed several significant insights.

4.2.1 General Perceptions

We first measured general perceptions of individual skill levels, threat level of attacks, and ideas on software piracy and self-efficacy levels of protecting against attacks. Figure 3 summarizes our findings. Our scores here are all on a 5-point Likert scale. From our results, we that computer skill (mean=3.2) is generally lower than mobile phone skill (mean=4.1), but they converge at the higher levels of education. Since this is survey data, we cannot say if there is a causal relationship, but these self-reported skill levels for computers and mobiles are both positively correlated with education level ($p<0.0001$), ($p=0.0020$) respectively.

We also find that feelings of 'dread', e.g. worrying about security (mean=4.0) and thinking that you could be a target for hackers (mean=3.6) are relatively high. Both measures of dread are positively correlated with self-reported skill with computers ($p<0.0001$), mobiles ($p<0.0001$), and education ($p=0.0039$), ($p=0.0125$). Meaning, despite increasing confidence in skill with technology and overall education level, worries about security and being attacked also increase. Also, respondent's overall concern of viruses being on computers was very high across the board (mean=4.4).

Interestingly, we found that respondents believing pirated software to be dangerous is somewhat high (mean=3.6) and positively correlated with education ($p=0.0003$), computer skill ($p=0.0305$), and mobile skill ($p=0.0216$). While the absolute numbers could be higher, this is positive finding because pirated software is a common vector for malware infection.

Finally, we found that confidence in protecting private information on computers is fairly high (mean=3.9). This confidence is not significantly correlated with education, but it is positively correlated with self-reported skill with computers and mobiles ($p<0.0001$). We explore the potential source of this confidence later in Section 4.2.5. Another interesting result was that we found that the general perception of mobile money transfer safety was high (mean=4.1) and was positively correlated to computer skill ($p=0.0001$) and mobile skill ($p=0.0006$), but not correlated to education. Mobile money transfers appear to be somehow outside of the categorically vulnerable set of Internet technologies. This may be yet another symptom of the conflation of network services discussed previously in Section 4.1.2.

4.2.2 Quality and Security

One finding from our interviews was that there was a strong correlation between perceived quality of websites and their perceived security or safety. This finding corroborates previous findings by Lindgaard [39] in the U.S. on judgments of trust being linked to appearance of webpages. When asked how users determined whether a website was safe, one respondent, for example, said:

Respondent: *Anytime I go on it, and it does not hesitate giving me information, that's why I think it's safe.*
Interviewer: *So because the information comes fast, you think it's safe?*
Respondent: *Yes.*

In the same vein, another expressed that those sites that return reliable information rather than false information are safe. Other reponses mentioned things like popups and advertisements indicating low quality and therefore lack of safety and sites with pornographic content being inherently unsafe. This is in contrast to other cues, such as the lack of SSL encryption typically indicated by a padlock icon in the browser or a website asking for more information than should be required to gain access to a site or service, neither of which were mentioned by even relatively expert users such as IT technicians. This result corroborates with previous works that find user assessment of trustworthiness often relies on cues not designed as security features [33] and that a majority of users ignore SSL warnings in a wide variety of conditions [51].

Various other measures of quality were expressed. Sites that send spam emails were unsafe; users said in response to this question that they stopped accessing websites if they became slow, etc. It is unclear whether these perceptions are due, as we suspect, to some sense that websites that seem well-made would naturally pay more attention to security in the same way that any well-manufactured object inspires confidence, or to some difference in the lexical range of the words "safe" and "trustworthy" in these contexts of which we are not aware. Evaluations were very subjective. One respondent identified unsafe websites as ones that "look mischievous". These findings are also consistent with findings by Wash that describe some users whose mental models dictate that they should only browse webpages from trustworthy sources [53]. We did not find any mention of the more sophisticated measures found by Wash during our interviews (e.g. disabling scripting, not clicking on attachments, or being careful downloading from websites).

4.2.3 Imputed Trustworthiness

In the same vein, many users commented that rather than trying to determine what sites were safe or not, they simply restricted their Web use to a handful of well-known sites such as Facebook, Google, Yahoo!, Wikipedia, and other sites that were recommended by friends, the referral by the crowd or by their friends thereby imputing some measure of trust. As a result, very few respondents said that they surfed the web by browsing - clicking through as they found links of interest in undirected exploration - rather, over 83% of interview respondents said that they went to specific sites or executed specific searches on trusted sites. One respondent mentioned that he never fills out online forms and when they are encountered he leaves the page.

The use of Google and Yahoo! raised a question for which we were unable to find a clear answer from our respondents: if search engines such as Google and Yahoo! are considered safe, are the links that they return in response to queries also considered safe as a result? Do the perceived brand quality or reliability of these search engines have a halo effect, or a social capital-like referrer effect on the returned pages, passing on imputed trust? Moreover, does this mean that users are abdicating the responsibility to understand whether pages are safe, relying upon these web properties to take care of that for them as described by Dourish [23]?

4.2.4 Security Measures

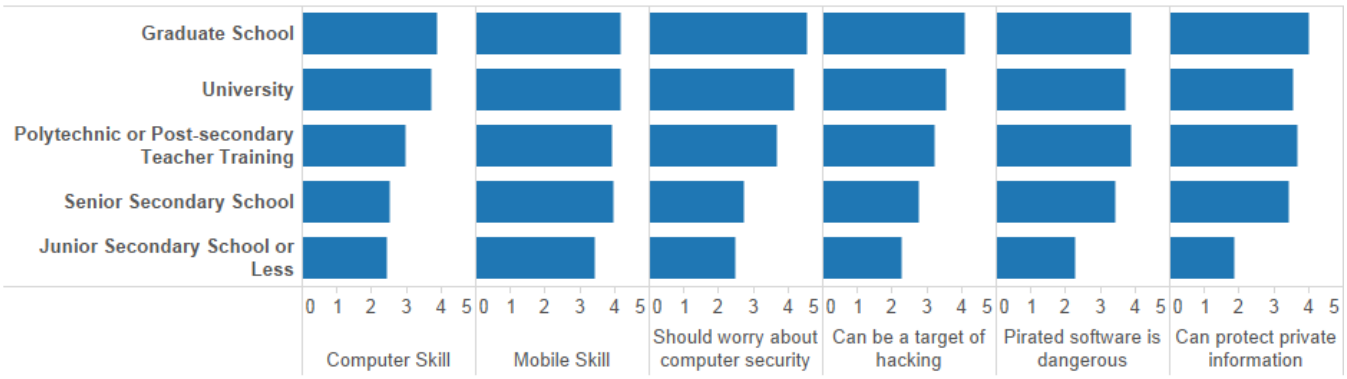


Figure 3: Self-reported skill levels and and perceptions of security and privacy on a 5-point Likert scale.

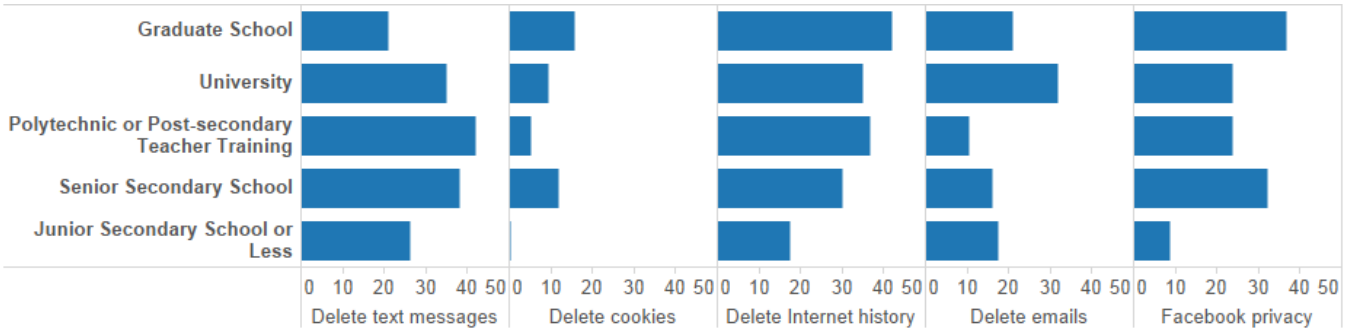


Figure 4: Security measures taken by percentage of respondents.

Measure	% Resp.	Education	Computer Skill	Mobile Skill
Delete texts	35.2%	0.13	0.67 **	0.59 **
Delete cookies	9.3%	0.30 **	0.02 .	0.02
Delete history	32.6%	0.05 *	0.07 * * *	0.06 **
Delete emails	21.2%	0.02	0.03 .	0.04 *
Facebook privacy	25.9%	0.02	0.06 **	0.05 **

Significance codes: 0 '* * *', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 '' Coefficients for correlation from a linear regression where values are on a 100-point scale vs a 5-point Likert scale for education levels.

Table 6: Correlations between defensive measures taken and education, computer skill, and mobile skill.

One of the primary research goals motivating our study was to determine what behaviors characterize the measures that people in these contexts took in order to protect their security and privacy online, and whether such measures were correct, commonly held, and adequate. We asked our respondents several survey questions on specific measures taken to defend against attacks. Figure 4 and Table 6 summarize our results. These results show that only up to 35.2% of respondents use even the most basic measures (deleting texts) to secure their private information. Other simple measures such as deleting history and emails follow close behind, but the instances of deleting cookies is far lower at 9.3% of respondents.

Surprisingly, 25.9% of respondents used Facebook privacy settings, which is high considering its complexity relative to simple deletions, but this may be due to the overall high level of Facebook use. From our results we find that users deleting Internet history is correlated to education and skill level. Deleting cookies, however, is only correlated to education level and computer skill level, but, unsurprisingly, not correlated to mobile skill level. We find that computer skill is correlated to performing all security and privacy measures. However, education is not correlated to deleting texts, deleting emails, or Facebook privacy settings. Figure 4 visually illustrates these trends.

During our interviews we directly asked interviewees how they stayed safe online. The most popular method by far was the use of a password, which we examine in greater depth in the following subsection. Other responses varied from nothing:

Interviewer: *What do you do to stay safe on the Internet?*
Respondent: *I don't do anything.*
Interviewer: *You don't do anything?*
Respondent: *Yeah.*

to relatively sophisticated measures including deleting cookies, deleting Internet browser history, private Googling (which we take to mean something akin to Incognito mode in Chrome), deleting chat history, logging out, rebooting the computer when done, not saving anything to desktop, restricting privacy settings on Facebook, avoiding unknown sites, and avoiding unknown people on social networking. We did not capture quantitatively in our surveys the prevalence of these more sophisticated measures other than deleting cookies and Internet history.

Through our interview data, we found that these measures were rarely used in any coherent regime, but were assembled ad hoc from information gathered from hearsay from various sources. This finding closely reflects previous work by Rader on stories acting as informal lessons about security [46]. Several respondents noted that they learned their safety measures from assistants at Internet cafés or had learned from friends, but only after they proactively requested help; this type of information does not appear to be proactively disseminated, according to our respondents.

Interviewer: *And how did you learn how to clear your history? Who taught you, or how did you know how to do it?*
...
Respondent: *That is the café assistant. I asked him 'What can I do so that people cannot gain access to my account?' And he tell me this is the way you can do it.*

It does appear that, among our respondents, there is a common distrust of shared computers (with strangers), which may be helping drive the adoption of connectivity via mobile phones and away from places like Internet cafés:

Interviewer: *Do you ever feel like it's unsafe to go on the Internet?*
Respondent: *Yeah, sometimes, sometimes I feel unsafe- especially when I go to the Internet café. There are people there who are also waiting for you to go so they also come. And they will be pressuring you to leave there so that they come.*
Interviewer: *So how is it unsafe?*
Respondent: *Maybe they can go to your history, the web history and then get access to your password, and then go into your accounts.*

and schools:

Interviewer: *So, what do you do to stay safe on the Internet?*
Respondent: *I browse at home most of the time, or at work, but I don't browse at school. Yeah, so at work I'm sure we're just using the work's, the office Internet, and then at home I have my own modem. So that's what I do.*

Aside from perceived quality, as mentioned earlier, users could not typically ascertain which sites were safe and which were not. Some relied upon software like antivirus programs, others explicitly claimed ignorance on the matter, a few were skeptical about security and felt that even commonly-used sites and services like Facebook and Skype were unsafe. One user expressed a sentiment that appeared to be widely held:

Respondent: *If it has a password, a place where you can put your password, only you can get access to it, and I know it's safe.*

4.2.5 Passwords

Passwords were the *de facto* gold standard for security among those interviewed. Of those asked about safety measures they took on the Internet, over 76% expressed that use of a password in one form or another was their only or primary means of staying safe; no doubt was expressed about the security of password mechanisms. Passwords were commonly recycled across all websites used, when asked "how often do you use the same password or PIN on different accounts", respondents responded with a mean score of 3.6 (moderately often). The distribution of rate of password reuse appears inverse normal (i.e. people either reuse their passwords often or never at all). 53% of respondents always or often use the same password for different accounts. 58% of respondents never changed their passwords, 22% changed their passwords once a year and 15% changed passwords once a month or less (5% did not respond to this question). Only a single person discussed password strength during the interviews, several respondents explicitly mentioned that they never changed passwords.

Passwords were considered effective so long as two measures were taken. The more commonly mentioned was memorization of the password (80% of respondents) as opposed to writing it down (24% of respondents):

Respondent 1: *[. . .] and then there's password. And, my password, I always memorize it so it's hard for you to get my password and access my stuff on the Internet as well.*
Respondent 2: *Normally I memorize my password, I always memorize on it.*
Respondent 3: *Yeah, I have a password [. . .] I keep it for memory.*
Respondent 4: *I've never changed a password [. . .] It's off head.*

Other respondents mentioned not sharing passwords except with a few trusted people such as family or close friends (10% of respondents).

An interesting and unforeseen effect of this implicit trust in passwords is that many users held sensitive personal information, including other passwords, in password-protected devices or services. Two examples of this were in email:

Interviewer: *How do you stay safe on the Internet?*
Respondent: *By keeping my informations in my email and then locking it up with my password.*

and on phones:

Respondent 1: *To avoid everything, I normally put passwords or PIN on my mobile phones. But apart from that, let's say if someone gets access to my, I wouldn't like them to see my financial information, maybe my personal photos or maybe my bank account details [. . .]*
Respondent 2: *I have account numbers on my phone, like my bank account number, I have it on my phone [. . .] I use a lot of password to block so that people might not see it.*

with the latter being far more common. Types of information held on phones included bank balances, bank account details, passwords for websites, medical and health information, and PIN codes. Of our respondents, 7% sent passwords to themselves via email and 7% did so through text messages. While this appears to be unsafe by security experts, considering the threats both perceived and real that face Ghanaian users this may in fact be a fairly rational practice.

4.2.6 Perceived Threat Model

It was clear from the responses and the types of approaches that respondents were using to stay safe during online activities that the mental model that users had of potential threats was significantly different than users in developed countries, perhaps more closely reflecting the actual threat model on the ground in their context.

Nearly all respondents expressed fears and to our respondents countermeasures such as passwords surrounded the human-computer interface. The most clear and present danger was from the person to the right or left. In other words, the context where respondents

accessed the Internet and threats from humans were either physically present or would be at some later time. As a result of this mental model, passwords were considered a strong safety measure so long as they were kept secret, as a human (or so they perceived) would find it impracticable to guess a password at random. Our findings here largely echo those by Klasnja *et al.* where relatively low user understanding of the underlying technology results in the dominance of a physical threat model [36].

This physical threat model is even further narrowed to people who make use of accounts that have not been logged out of, which is potentially the most common attack vector for this user base. Interestingly, despite the focus on threats from people physically nearby, no mention was made of shoulder surfing, keyloggers, or other slightly more sophisticated methods of local attack at the man/machine boundary layer.

Users displayed high confidence in the security of systems that they had logged into, as evidenced by the use of email for storage of sensitive information. When pressed on the possibility, for instance, that someone could intercept chat information, users were not concerned:

Interviewer: *Do you think anybody could take your conversation and do something with it?*
Respondent 1: *No, because we chat alone. So no one can hear any information about us.*

Interviewer: *Do you ever worry that your chats are being saved somewhere, and someone's using the information for something else?*
Respondent 2: *No, I don't think somebody can use my information.*

Especially noteworthy to us is the first response above; the respondent goes on to explicitly state that no one can get that information unless they get into his email, and that's not possible because he always logs out, clears history, and reboots the computer. Again, the attack surface, in the respondent's mind, was limited to the particular physical terminal that he used - the network beyond that terminal represented a safe zone. This appears to be a very commonly held belief, that while the network may go down between the terminal and the site, no other danger exists in the network; that it is effectively a direct link between the terminal and the various sites, and that danger must come either from the site accessed or at the terminal; that no danger can be interjected between the two. Again, our results here corroborate very closely with findings by Klasnja *et al.* [36] that show how users often have no awareness of data visibility when interacting with a remote web server through a network.

Other aspects of the threat model were unusual to our minds as well - users had implicit trust that their phones would not be compromised if they had passwords - this despite many users specifically mentioning that the reasons that they chose the phones that they did was because there were many phone shops that could repair them. These same repair shops could, of course, also reset the passwords and access whatever is inside, something that did not appear to occur to any of our respondents. Furthermore, from our interviews respondents appeared to understand the difference between phone passwords and a "SIM passwords" (SIM PIN). We did not have quantitative results for proportion of users using each kind of password, but in our interviews we found a predominance of phone passwords being used. It is possible that SIM PINs are only used when necessary in cases such as repair shops or users simply do not worry as much about their contacts being stolen.

Further, aside from scant mentions of antivirus software, the topic of viruses and malware never arose, despite having among the highest infection rates in Africa [14]. It is particularly unclear whether any participants were aware of the various forms of malware that capture passwords and other information that is entered into computers and how that may have affected their opinions of the use of passwords.

4.2.7 Fear of and Experience of Hacking

In our quantitative findings we discovered a high level of dread related to hacking and being targeted. However, during interviews respondents did not consider hacking an immediate threat. While many had heard of hacking, few had a clear idea of what it entailed or what possible repercussions could occur. As with the threat model described above, people's idea of the danger of hacking was mostly limited to those threats in the immediate vicinity.

A select few respondents had direct knowledge of hacking as victims, but only one displayed understanding that transcended guesswork:

Respondent: *Yes, one of my accounts has been hacked. [. . .] It's like PayPal. So they hacked it, immediately I transferred money to like, under a few seconds they took the money. I transferred 200 Cedis [(20 USD at the time)] into it, and someone else from nowhere took the money. They started tracking the IPS [sic] address and they were like, it's in India or something, but I just forgot about it. And since then I've never done anything online transaction.*

Other firsthand victims of hacking had much more benign stories, mostly of having their Facebook accounts broken into, or their email accounts broken into and passwords changed to lock them out. Secondhand stories, included friends whose email accounts had been hacked and the accounts used to send email to friends asking for money, a friend who had posted his bank account details online without a password and whose account was promptly emptied, someone who had all his money stolen by someone in France, someone whose Mastercard was hacked, and various other perfidy.

The concept and scope of hacking is vaguely defined. The term 'hacking' may include activities such as phishing, scams, spam, etc. Most of our respondents use hacking as an umbrella term rather than more specific terminological distinctions. To our respondents hacking included scams (including 419 scams) and phishing. One user who was 'hacked' had responded to a phishing text and had his MTN (a major GSM cell operator in Ghana [9]) phone credit balance stolen because he provided his PIN.

The potential consequences of being hacked tended to gravitate around three potential outcomes. First, several respondents indicated that a hacker who gained access to their online accounts would ask their friends for money:

Respondent: *Most hackers will send mails asking for money. So, maybe ask for money from my friends. That's what most hackers do.*

A second possible outcome is theft of personal information, again, for the purpose of stealing money:

Respondent: *[. . .] and get sensitive information like my bank details, my personal information, and use it against me.*

A third major possibility expressed was the nuisance of being locked out of their own accounts, as some other respondents had experienced firsthand. Others mentioned that hackers might blackmail them, or implicate them in a hacking attack on another per-

son, use their account to send out spam messages, do damage to their work, etc., but the most common fear is the direct loss of money. In relation to the mental models about hackers as described by Wash [53], most of our respondents' mental models could be captured by the "Burglar" folk model: the identity of the hacker is "some criminal whose reason for break-ins is to look for financial and personal information and possibly harm the computer or expose personal information opportunistically".

The way people are hacked was not made clear to us. As is consistent with our prior observations about attitudes towards passwords and threat models, in general people blamed hacking attacks on lack of passwords or people having given out their passwords.

5. DISCUSSION

We have found that there are substantial security gaps (according to common 'best practice' security advice) in the way online services are used by urban Ghanaians. Online threats are global, but perception of threats, in general, are very localized. Informal lessons result in a patchwork of ad hoc mechanisms being used to secure personal information. Password use, deleting messages, emails, and browser history are currently the key mechanisms for protecting against hackers. Network technology is mentally construed as being a black box; what goes on behind the screen is not part of the mental threat model. The conflation of services and agnosticism to the device or software application being used are also suggestive of this mental model.

We also found that, like in rich countries, people's perceptions of trustworthiness are also predominantly ad hoc and from the perspective of the immediately visually apparent. E.g. appearance, lack of popups, loading speed, specific safe websites, etc. People's confidence in their ability to defend themselves against security threats is similarly based on the apparent. We found that the most common defense is to depend on passwords and memorization of passwords. Unfortunately, passwords are rarely changed by most and stored in an unsafe manner and often reused. We found there is a strong worry about security and of being hacked, possibly due to the unknown nature of hacking, but despite this concern, respondents reported feeling that they were able to defend themselves despite passwords often being the only line of defense. This confidence is likely due to the low incidence of hacking. Finally, we found that the concept of hacking being typically confined to stories and conceptions of private information being stolen and monetary loss.

Despite this possibly bleak picture of Internet security in Ghana, given the low incidence of local cybercrime, the mental threat model and existing practices actually appear largely adequate for the time being. Unlike in rich countries where users are largely ignoring onerous security advice [31], we found that some users actually go to great lengths to protect their security and privacy even if the way they do so is imperfect (clearing history, deleting messages, etc). Social engineering by 419 scams and phishing in spam are relatively well known to our respondents and are mostly captured by the existing mental model and countermeasures. Other types of hacking such as large-scale data theft and botnet infections that fall outside of the existing mental model have not yet occurred likely due to the present lack of profit to be made when compared to targets in rich countries.

While the existing defensive measures may be sufficient and even appropriate for the actual threats on the ground at present, given the continued trends it is unlikely that this will continue to be the case. As network bandwidth increases along with penetration, the restriction of Internet use primarily to a few popular sites, is unlikely to hold, and Ghanaian Internet users will become exposed to the wider array of Internet-based threats including, but not limited to, malware, phishing, and various illegitimate sites. Of special concern is the fact that as bandwidth increases and costs come down, use of the Internet (as opposed to burned CDs as are currently the more popular option) for the acquisition of pirated software, a popular vector for malware, will likely increase. This is not currently a problem for devices that are often disconnected or have low bandwidth, but as connectivity improves these devices may be more attractive to attackers. Already some interview respondents mention using the Internet to visit "warez" sites to download software.

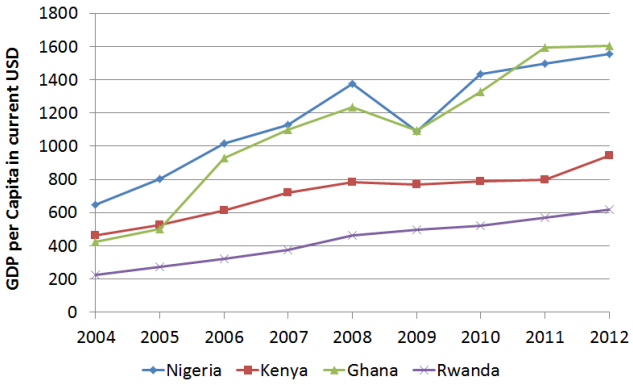


Figure 5: GDP per capita in current USD, from [2]

Further compounding the near-term threat is the general trend towards affluence in many sub-Saharan nations such as Ghana, Nigeria, Kenya, and Rwanda, a trend clearly seen in Figure 5. As users, on average, become wealthier, they will naturally become riper targets for online exploitation of various kinds aimed at appropriating that wealth. Finally, the promulgation of mobile financial, health, and governmental services in these developing contexts without commensurate security precautions is of concern.

5.1 User Education and Threat Mitigation

To mitigate the confluence of these trends, all of which will tend to reduce the security of the average user of the Internet in Ghana, steps could be taken proactively. One possibility is to educate users on the the reality of the types of threats they may face on the wider web. Rather than the ad hoc self-education our respondents reported, more education and resources could be delivered to users of the Internet. Unfortunately, conventional security awareness programs are unlikely to completely solve the problem when security advice continues to grow in complexity and following this advice has been shown to have unclear benefits [31]. Instead, targeted security advice specific to particular applications and services may be more easily followed if easier to follow. Much as health information is delivered in increasingly clever ways, information about avoiding hazards online might be delivered packaged with the service being used. For example, on SMS applications security advice could be sent as informational SMSs as part of the service or for mobile data plans the mobile-operator could give advice.

Another possible focus of education is for the common user to be made aware of the nature of the Internet. Basic concepts such as there being an ungoverned expanse between the user's terminal and the site they are trying to access, the importance of the use of technologies like SSL to prevent man-in-the-middle attacks, traffic sniffing, etc. could prove to be eye- opening and might cause a change in security-related behaviors. We have not quantitatively studied the prevalence of the notion that passwords are

impregnable, but if this is indeed the case then educating users or demonstrating the fallibility of passwords e.g. using John the Ripper [8] might prove enlightening. Similarly, warnings that passwords on smartphones and feature phones alike can be bypassed and, as such, that phone handsets do not serve as secure repositories, would likely be of help.

Given the current preference for mobile phones and passwords, two-factor authentication as recently employed by Facebook, Google, and Yahoo [3, 6, 12] may be more appropriate if the second authentication factor were not tied to the mobile phone. In addition to this it may, at least in the near term, be advisable to set up ISP-based blocking on sites known to carry malware or questionable content. However, this type of regulation creates censorship challenges and is also unlikely to help people who will specifically be looking for pirated software or illegal music or media downloads.

The ad hoc nature of communication of security information may, alternatively be leveraged through the use of social networks, making use of social capital within social graphs to improve uptake of informal security stories and security advice [17].

5.2 Avenues for Further Research

Our findings thus far suggest avenues for further research. Evaluating how users in this context develop their mental threat models could prove to be fruitful despite their fundamental complexity - to what extent these models are based on hearsay through their social graph, personal experience, news, and other sources is certainly worthy of deeper investigation. Also interesting would be an analysis of how imputed trustworthiness works - whether sites are perceived to lend legitimacy to sites they link to by default - and whether this can be modeled in the same way as social capital flows through social graphs. Google's PageRank already incorporates imputed trustworthiness to a degree - pages linked from reliable pages are considered more reliable - so these types of assumptions, depending on search terms, may not be far off.

Also worthwhile would be an effort to front-run the inevitable increase in hacking and establish certain baseline attitudes and practices, and evaluate how these evolve over time as this increase takes place. It is also unclear whether greater use of mobile phones rather than computers to access the Internet result in less worry about viruses and malware.

Finally, new mechanisms for usable security and privacy designed to be appropriate for these developing region contexts could have a big impact as existing mechanisms transplanted along with the default technologies do not appear to be widely adopted. There may be interesting opportunities for novel solutions based due to mobile phones being the primary means of access to services.

6. CONCLUSION

This paper makes two main contributions. First, this is the first study to our knowledge to focus on exploring security perceptions and practice in a developing country context. We have examined technology users throughout Ghana to comprehensively understand the technology landscape and people's perceptions regarding security. Second, we examined the security measures that people take to protect themselves online. We found in our survey corresponding to 193 participants that the characteristic attitudes include: 1) reliance and trust in password systems, 2) vague understanding of how networked systems work and therefore what factors constitute realistic threat models resulting in an asymmetric focus on local threats, 3) a conflation of perceptions of quality and perceptions of security, consistent with existing research, and 4) various observations on security-related behaviors, Internet and social network usage patterns, and other miscellany.

Interestingly, the physical threat models and lack of understanding of how networked systems work are very similar to previous findings in rich countries. The ad hoc acquisition of security knowledge is also similar to previous findings. The difference in Ghana is that the low incidence of local cybercrime makes these existing threat models and practices relatively adequate for the time being. Some would argue that this is the case even in rich countries, but given the continued trends in Internet penetration, income, and dependence on the network for basic services we do feel that this is a risky proposition.

It is yet unclear to what extent the users we are interacting with can serve as representative of users elsewhere in sub-Saharan Africa or the developing world as a whole, but we hope our contributions are able to help characterize the overall shape of security in the developing world and provide a starting point for discussion and research.

7. ACKNOWLEDGMENTS

We thank the respondents who participated in the study and our reviewers for their helpful feedback. We would also like to thank our shepherd Joseph Bonneau for his help improving this paper.

8. REFERENCES

[1] CIA World Factbook. **https://www.cia.gov/library/publications/the-world-factbook/fields/2177.html**.

[2] Data | The World Bank. **http://data.worldbank.org/**.

[3] Facebook Login Approvals. **http://www.facebook.com/note.php?note_id=10150172618258920**.

[4] Facebook Reports Second Quarter 2013 Results. **http://investor.fb.com/releasedetail.cfm?ReleaseID=780093**.

[5] Ghana Economic Outlook. **http://www.afdb.org/en/countries/west-africa/ghana/ghana-economic-outlook/**.

[6] Google 2-Step Verification. **https://www.google.com/landing/2step/**.

[7] Hey! My friend's account was hacked! **http://blogs.windows.com/windows_live/b/windowslive/archive/2011/07/14/hey-my-friend-s-account-was-hacked.aspx**.

[8] John the Ripper password cracker - Openwall. **http://www.openwall.com/john/**.

[9] MTN. **https://www.mtn.co.za/Pages/Home.aspx**.

[10] The Quiet Mobile Giant: With 300M Active Users, WhatsApp Adds Voice Messaging. **http://allthingsd.com/20130806/the-quiet-mobile-giant-with-300m-active-users-whatsapp-adds-voice/**.

[11] WhatsApp :: Home. **http://www.whatsapp.com**.

[12] Yahoo 2-Step Verification. **https://edit.yahoo.com/commchannel/sec_chal_manage**.

[13] Measuring the Information Society. **http://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2012/MIS2012_without_Annex_4.pdf**, 2012.

[14] Global Virus Map. **http://home.mcafee.com/virusinfo/global-virus-map**, 2013.

[15] Bagchi, K., Kirs, P., and Cerveny, R. Global software piracy: can economic factors alone explain the trend?

Communications of the ACM 49, 6 (2006), 70–76.

[16] Ben-David, Y., Hasan, S., Pal, J., Vallentin, M., Panjwani, S., Gutheim, P., Chen, J., and Brewer, E. A. Computing security in the developing world: A case for multidisciplinary research. In *Proceedings of the 5th ACM workshop on Networked systems for developing regions*, ACM (2011), 39–44.

[17] Besmer, A., Watson, J., and Lipford, H. R. The impact of social navigation on privacy policy configuration. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, ACM (2010), 7.

[18] Brewer, E., Demmer, M., Du, B., Ho, M., Kam, M., Nedevschi, S., Pal, J., Patra, R., Surana, S., and Fall, K. The case for technology in developing regions. *Computer 38*, 6 (2005), 25–38.

[19] Burrell, J. *Invisible users: Youth in the Internet cafés of urban Ghana*. Mit Press, 2012.

[20] Burrell, J. Technology hype versus enduring uses: A longitudinal study of internet use among early adopters in an african city. *First Monday 17*, 6 (2012).

[21] Cyr, D. Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty. *Journal of Management Information Systems 24*, 4 (2008), 47–72.

[22] Cyr, D., Head, M., and Larios, H. Colour appeal in website design within and across cultures: A multi-method evaluation. *International Journal of Human-Computer Studies 68*, 1 (2010), 1–21.

[23] Dourish, P., Grinter, R. E., De La Flor, J. D., and Joseph, M. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing 8*, 6 (2004), 391–401.

[24] Egelman, S., Cranor, L. F., and Hong, J. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2008), 1065–1074.

[25] Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., and Herley, C. Does my password go up to eleven?: the impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 2379–2388.

[26] Everard, A., and Galletta, D. F. How presentation flaws affect perceived site quality, trust, and intention to purchase from an online store. *Journal of Management Information Systems 22*, 3 (2006), 56–95.

[27] Fair, J. E., Tully, M., Ekdale, B., and Asante, R. K. Crafting lifestyles in urban africa: Young ghanaians in the world of online friendship. *Africa Today 55*, 4 (2009), 29–49.

[28] Fischer, P., Kastenmüller, A., Greitemeyer, T., Fischer, J., Frey, D., and Crelley, D. Threat and selective exposure: The moderating role of threat and decision context on confirmatory information search after decisions. *Journal of Experimental Psychology: General 140*, 1 (2011), 51.

[29] Florêncio, D., Herley, C., and Coskun, B. Do strong web passwords accomplish anything. *Proc. Usenix Hot Topics in Security* (2007).

[30] Grinter, R. E., Edwards, W. K., Newman, M. W., and Ducheneaut, N. The work to make a home network work. In *ECSCW 2005*, Springer (2005), 469–488.

[31] Herley, C. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, ACM (2009), 133–144.

[32] Herley, C., and Van Oorschot, P. A research agenda acknowledging the persistence of passwords. *Security & Privacy, IEEE 10*, 1 (2012), 28–36.

[33] Jakobsson, M., Tsow, A., Shah, A., Blevis, E., and Lim, Y.-K. What instills trust? a qualitative study of phishing. In *Financial Cryptography and Data Security*. Springer, 2007, 356–361.

[34] Karaganis, J. *Media piracy in emerging economies*. Lulu. com, 2011.

[35] Kim, J., and Moon, J. Y. Designing towards emotional usability in customer interfaces.trustworthiness of cyber-banking system interfaces. *Interacting with computers 10*, 1 (1998), 1–29.

[36] Klasnja, P., Consolvo, S., Jung, J., Greenstein, B. M., LeGrand, L., Powledge, P., and Wetherall, D. When i am on wi-fi, i am fearless: privacy concerns & practices in eeryday wi-fi use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2009), 1993–2002.

[37] Kowitz, B., and Cranor, L. Peripheral privacy notifications for wireless networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, ACM (2005), 90–96.

[38] Kuo, C., Romanosky, S., and Cranor, L. F. Human selection of mnemonic phrase-based passwords. In *Proceedings of the second symposium on Usable privacy and security*, ACM (2006), 67–78.

[39] Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., and Noonan, P. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction (TOCHI) 18*, 1 (2011), 1.

[40] Mas, I., and Morawczynski, O. Designing mobile money services lessons from m-pesa. *Innovations 4*, 2 (2009), 77–91.

[41] Morris, R., and Thompson, K. Password security: A case history. *Communications of the ACM 22*, 11 (1979), 594–597.

[42] Norman, D. A. The way i see it when security gets in the way. *interactions 16*, 6 (2009), 60–63.

[43] Paik, M. Gotta catch'em all!: innoculous: enabling epidemiology of computer viruses in the developing world. In *Proceedings of the 5th ACM workshop on Networked systems for developing regions*, ACM (2011), 51–56.

[44] Pal, J., Nedevschi, S., Patra, R. K., and Brewer, E. A. A multidisciplinary approach to open access village telecenter initiatives: The case of akshaya. *E-Learning 3*, 3 (2006), 291–316.

[45] Panjwani, S. Towards end-to-end security in branchless banking. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, ACM (2011), 28–33.

[46] Rader, E., Wash, R., and Brooks, B. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, ACM (2012), 6.

[47] Schmidt, M. B., Johnston, A. C., Arnett, K. P., Chen, J. Q., and Li, S. A cross-cultural comparison of us and chinese computer security awareness. *Journal of Global Information Management (JGIM) 16*, 2 (2008), 91–103.

[48] Shay, R., Komanduri, S., Kelley, P. G., Leon, P. G., Mazurek, M. L., Bauer, L., Christin, N., and Cranor, L. F. Encountering

stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, ACM (2010), 2.

[49] Smyth, T. N., and Best, M. L. Tweet to trust: social media and elections in west africa. In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers-Volume 1*, ACM (2013), 133–141.

[50] Sun, S.-T., Pospisil, E., Muslukhov, I., Dindar, N., Hawkey, K., and Beznosov, K. What makes users refuse web single sign-on?: an empirical investigation of openid. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, ACM (2011), 4.

[51] Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., and Cranor, L. F. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX Security Symposium* (2009), 399–416.

[52] Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., Passaro, T., Shay, R., Vidas, T., Bauer, L., et al. How does your password measure up? the effect of strength meters on password creation. In *Proc. USENIX Security* (2012).

[53] Wash, R. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, ACM (2010), 11.

[54] Wyche, S. P., Forte, A., and Yardi Schoenebeck, S. Hustling online: understanding consolidated facebook use in an informal settlement in nairobi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 2823–2832.

[55] Wyche, S. P., Schoenebeck, S. Y., and Forte, A. Facebook is a luxury: An exploratory study of social media use in rural kenya. In *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM (2013), 33–44.

Appendix A: Interview Questions

Talking points for interviews, number of ∗ indicates priority.

Security Practices, Attitudes, and Anecdotes

∗∗∗**1.** Have any of your accounts ever been hacked or do you know anyone who has had an account hacked?

∗ ∗ ∗**2.** What do you do to stay safe on the Internet?

∗ ∗ ∗**3.** How often do you use a pen-drive?

∗∗**4.** Is there any personal information, or anything you wouldn.t want other people to see on your phone?

∗∗**5.** If someone hacked your email, what other things could he do with your email account? (do you use the same email/password for other services, etc.)

Internet

∗ ∗ ∗**6.** How do you tell which web pages are safe or trustworthy and which are not?

∗ ∗ ∗**7.** [Ask what their email address is, check what information is public on their G+ or FB profile - if they have Twitter or equivalent, check visibility of their stream]

∗∗**8.** Do you ever search for your own name online?

9 What websites do you regularly visit?

∗∗**10.** Do you spend a lot of time on social network sites like Facebook, Google+, or Twitter?

∗**11.** When you use the Internet, do you usually go online for something specific (score of a football match, today.s news, information about jobs) or do you .browse. by clicking through from page to page?

12 How would your life be different if you couldn.t use the Internet?

13 Do you use email or SMS more?

Mobile Phones

∗∗**14.** Can you show me the kinds of things you do using your mobile phone?

∗∗**15.** Does your mobile phone have a password? If so, is the password for the phone or for the SIM?

∗**16.** Why did you choose the mobile phone you chose?

Appendix B: Intermediate Data

Resp. ID	What do you do on the Internet	Staying safe on the Internet	Social networking	...
53	search for engineering, design, business research	privacy settings e.g. facebook	yes	...
54	mail, jobs, fb, news, games, entertainment	passwords on documents	yes	...
55	research	don't open certain sites	no	...
56	company, contacts, work	don't keep personal info, email is encrypted	seldom	...
57	interact with friends and colleagues			...
...	...	...	...	...

Table 7: A fragment of the data matrix for analyzing interview data. This matrix includes the characteristic behaviors and comments from interviews.

The Effect of Social Influence on Security Sensitivity

Sauvik Das, Tiffany Hyun-Jin Kim, Laura A. Dabbish, Jason I. Hong
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
sauvik@cmu.edu, hyunjin@cmu.edu, dabbish@cmu.edu, jasonh@cs.cmu.edu

ABSTRACT

Despite an impressive effort at raising the general populace's *security sensitivity*—the awareness of, motivation to use, and knowledge of how to use security and privacy tools—much security advice is ignored and many security tools remain underutilized. Part of the problem may be that we do not yet understand the social processes underlying people's decisions to (1) disseminate information about security and privacy and (2) actually modify their security behaviors (e.g., adopt a new security tool or practice). To that end, we report on a retrospective interview study examining the role of social influence—or, our ability to affect the behaviors and perceptions of others with our own words and actions—in people's decisions to change their security behaviors, as well as the nature of and reasons for their discussions about security. We found that social processes played a major role in a large number of privacy and security-related behavior changes reported by our sample, probably because these processes were effective at raising security sensitivity. We also found that conversations about security were most often driven by the desire to warn or protect others from immediate novel threats observed or experienced, or to gather information about solving an experienced problem. Furthermore, the *observability* of security feature usage was a key enabler of socially triggered behavior change—both in encouraging the spread of positive behaviors and in discouraging negative behaviors.

1. INTRODUCTION

There are many reasons why security advice is often ignored and many security tools are left unutilized [17]. Some prior work suggests that many believe they are in no danger of experiencing a security breach [1] and are *unaware* of both threats and the security tools available to protect against those threats. Other work suggests that many choose not to use security tools and follow security advice because doing so is often antagonistic towards the immediate goal of end users—a complex password that usually requires three attempts to get right *prevents* a user from doing what she actually wants to do: e.g., authenticating into social media. Herley further argues it may even be economically *rational* for users to ignore security advice, as the expected cost, in time, of a lifetime of following security advice might actually be higher than the expected loss a user would suffer if his account actually was compromised [17]. Thus, many people lack the *motivation* to behave securely. Still others suggest that security tools are simply too difficult to use [26,34], so many people do not have the *knowledge* required to operate them. Taken together, it appears that the lack of what we call *security sensitivity*—the awareness of, motivation to use, and knowledge of how to use security and privacy tools—is a large barrier to increasing the uptake of security tools and the following of security advice.

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9-11, 2014, Menlo Park, CA.

Prior work has looked at improving all parts of the security sensitivity stack—for example, through games for security education [28], browser extensions to make users more aware of phish [35], more effective user interfaces for security tools [19], and faster or simpler ways to authenticate users [31]. Security sensitivity, nevertheless, remains low.

We argue that part of the problem is that we do not yet understand the *social* processes underlying people's decisions to communicate about security and adopt security tools. In other words, security behaviors—as any human behavior—should be viewed within the context of a social system. Indeed, social psychology and sociology literature illustrates that social influence, or our ability to affect other people's perceptions and behaviors with our words and actions [6], plays a central role in how people behave—even specifically in changing their behavior or adopting a new technology or idea [6,25]. Rogers' highly influential *diffusion of innovations* work, for example, has shown that social influence drives technology adoption [25]. Social processes, thus, should undoubtedly affect a user's decision to follow security advice or adopt a security tool.

Nevertheless, the effect of social influence on decisions and communications about security and privacy remains understudied. Indeed, we do not yet know *how* social influence affects behavior change with regards to security and privacy. Likewise, we know little about the nature of conversations about security and privacy, through which this influence should occur. Understanding how social influence affects security related behavior change and communication could improve our understanding of why security sensitivity remains low, and could help inform the design of social interventions that can raise security sensitivity. To that end, we report on a retrospective interview study aimed at investigating the following research questions:

Q1: What role does social influence play in an individual's decisions to use, discontinue use, and explore security tools and privacy settings?

Q2: Under what circumstances do people communicate about security and privacy?

In our interviews, we probed participants about their experiences with regards to mobile phone authentication, mobile application installation and uninstallation, and social media privacy settings. We also asked participants to recall specific conversations they had about cybersecurity and online privacy.

Our findings suggest social processes played a major role in a large number of privacy and security related behavior changes reported by our interviewees, probably because these processes were effective at raising all points of the security sensitivity

stack—awareness, motivation and knowledge. However, different triggers for socially driven behavior change varied in the extent to which they raised awareness, motivation and knowledge about security tools and behaviors. In addition, conversations our participants had about security and privacy were most often instigated by the desire to (1) warn or protect others from immediate or novel threats observed or experienced, and (2) to gather information about solving an immediate problem. One particularly salient theme that arose from our interviews is that the *observability* of security feature usage was a key enabler of socially triggered behavior change and conversation—in encouraging the spread of positive behaviors, discouraging negative behaviors, and getting participants to talk about security. Taken together, our results suggest that: (1) there is a substantial and often overlooked social process that strongly affects security-related behavior change, and (2) in order to maximally raise security sensitivity, security and privacy tool usage should be more observable and amenable to conversation.

2. RELATED WORK

2.1 Security Sensitivity

Prior work in usable privacy and security alludes to three reasons underlying why much security advice is ignored and many security features remain unused: lack of awareness, motivation, and knowledge. First, many users lack the *awareness* of security threats and the tools available to protect themselves against those threats. For example, a study by Adams and Sasse found that insufficient awareness of security issues caused users to construct their own model of security threats that are often incorrect, resulting in security breaches [2]. Stanton et al. found that a lack of awareness of basic security principles even influenced experts to make naïve security mistakes, such as using a social security number as a password [30]. Users who are unaware of a threat cannot take measures to avoid the threat, and users who are not cognizant of the tools available to protect themselves from these threats cannot use those tools to actively defend themselves.

Second, users—even those who are aware of security and privacy threats and the preventive tools that combat those threats—often lack the *motivation* to utilize security features to protect themselves [2,12]. The lack of motivation to use security features is not entirely surprising, as stringent security measures are often antagonistic towards the specific goal of the end user at any given moment [10,26]. For example, while a user might want to access her Facebook, a complex password that usually requires three attempts to get right *prevents* her from accessing Facebook for an intolerable amount of time [11]. Thus, users often reject the use of security and privacy tools when they expect or experience them to be weighty [2,14,18,26]—and security features are often weighty.

Furthermore, many security threats remain only an abstract threat to most individuals [2,16,24]: Bob may know, conceptually, that there are security risks to using the same simple password across accounts, but does not believe that he is, himself, in danger of experiencing a security breach. Additionally, this perspective may be economically rational, as the expected cost, in time, of a lifetime of following security advice might actually be higher than the expected loss a user would suffer if his account actually was compromised [17]. Finally, the benefits of security features are often invisible, as users are often not cognizant of the *absence* of a breach that otherwise would have occurred without the use of a security or privacy tool. In all, it is unsurprising that many users lack the motivation to explicitly use security tools: to do so would mean to incur a frustrating complication to everyday interactions in order to prevent an unlikely threat with little way to know whether the security tool was actually effective.

Third, security tools are often too complex to operate for even aware and motivated end-users, suggesting that users lack the *knowledge* to actually utilize security tools [34]. Indeed, there is a wide gulf of execution for most security features for most users. For example, many users cannot distinguish legitimate vs. fraudulent URLs, nor forged vs. legitimate email headers [8]. Also, a study revealed how security features in Windows XP, Internet Explorer, Outlook Express, and Word applications are difficult for users [13]. And, Wash found that many people hold "folk" models of computer security that are often misguided, and use these incorrect models to justify ignoring security advice [32].

In sum, prior work in usable privacy and security suggests that there are at least three large obstacles inhibiting the widespread use of security and privacy tools: the *awareness* of security threats and tools, the *motivation* to use security tools, and the *knowledge* of how to use security tools. We refer to this layered stack as *security sensitivity* for ease of discussion, as it encapsulates how likely a user is to seek information about and use security tools. Note, however, that the concept is not necessarily novel, as prior work has alluded to such a stack in security specifically [12], and in the adoption of technology more generally [7,25].

2.2 Social Influence and Security Sensitivity

In his seminal work on the diffusion of innovations, Rogers claimed that new technology gets widely adopted through a process by which it is communicated through members of a social network [25]. Rogers argues that primarily *subjective perceptions*, not scientific or empirical fact, get communicated through social channels, and that these perceptions are key to the success of an innovation. He further outlines that preventative innovations—or innovations, like security and privacy tools, that prevent undesirable outcomes from happening in the future—typically have low adoption rates, probably because of their lack of *observability*, or the invisibility of their use and benefits.

Other work in cognitive psychology has looked at the psychological mechanisms underlying social influence. For example, lots of prior work has demonstrated the potency of the concept of "social proof"—or our tendency to look to others for examples of how to act in uncertain circumstances [5,6]. For example, Milgram, Bickman, and Berkowitz [22] demonstrated the social proof principle when they showed that simply getting a small crowd of people—the more, the better—to look up at the sky on a busy sidewalk caused others to do the same.

Still other work has shown how social interventions can be powerfully effective at driving human behavior: for example, at reducing household energy consumption by showing people their neighbors' reduced energy consumption [27], reducing hotel guests' wasteful use of towels by showing them previous patrons chose to be less wasteful [15], and even in eliminating young children's phobia of dogs by showing them film clips of other children playing with dogs [3].

Taken together, the background literature suggests that social influence strongly affects people's behaviors and decisions; likely, also their security-related behaviors and decisions. And, indeed, prior work *has* alluded to the importance of social processes in raising security sensitivity. For example, DiGioia and Dourish [9] suggested that "social navigation"—or people's inclination to

	Age	Gender	Race	Occupation
P1	28	Male	Black	Customer Service
P2	22	Female	Asian	Unemployed
P3	22	Female	Black	Student
P4	22	Male	Black	Student
P5	27	Female	Asian	Unemployed
P6	29	Male	White	Programmer
P7	54	Female	White	Admin. Assistant
P8	31	Male	Indian	Unemployed
P9	30	Male	White	Software Developer
P10	37	Male	White	Graphic Designer
P11	54	Male	Black	Chef
P12	20	Female	Black	Student
P13	24	Female	Indian	Graduate Student
P14	25	Male	Indian	Graduate Student
P15	21	Male	Indian	Graduate Student
P16	22	Male	Indian	Graduate Student
P17	34	Female	Asian	Unemployed
P18	20	Male	Black	Student
P19	20	Male	White	Student

Table 1. Participant demographics.

look for cues on how to act—can be used to raise users' security sensitivity by showing them other users' actions in context. Rader et al.'s study on stories as informal lessons about security suggests that storytelling increases awareness of and motivation to guard against security threats [23]. In addition, Singh et al. outlined the common practice of sharing passwords and PINs [29]. On the other hand, Gaw et al. [14] found that many people believed that use of security features was an indication of paranoia, unless the user had an obvious reason for doing so. If there *is* a stigma of paranoia attached to using security features, then it is possible that, under some circumstances, social influence can work *against* security sensitivity (e.g., "only paranoid people encrypt their e-mail, and I'm not paranoid").

Nevertheless, the background literature on the social dimensions of security and privacy remains surprisingly thin. To our knowledge, little work has specifically looked at how social influence affects security sensitivity, and, in turn, enacts behavior changes related to privacy and security, or how people generally communicate about security and privacy (outside of Rader et al.'s study on security storytelling [23]). Yet, understanding how social influence affects security related behavior change and communication could improve our understanding of why security sensitivity remains as low as it is, and could even help inform the design of social interventions that raise security sensitivity. To that end, we look specifically at the social dimensions of security related behavior changes and communications in this work.

3. Methodology

3.1 Semi-Structured Interview Methodology

We constructed an IRB approved semi-structured interview protocol to probe participants about recent security related behavior changes and conversations. We elected a semi-structured approach so that we could concretize the discussion by directing participants' memories towards *changes* in behavior or *specific* instances of communication, while still allowing participants the flexibility to expand on the their undoubtedly unique experiences. Our interview protocol probed participants about recent changes in (1) *mobile authentication*, or whether and why participants enabled, disabled, or changed authentication on their smartphones (e.g., from PIN to Password); (2) *application installation and uninstallation*, or whether and why participants decided to uninstall or halt installing applications because of privacy and security concerns; and, (3) *online privacy settings in social media*, or whether and why participants changed their privacy settings on the social media platform they most commonly used. We chose to explore three categories to uncover general trends across different types of security tools, and we chose these three categories specifically because they represented a broad range of behaviors representative of common security and privacy decisions made by just about all people fairly regularly.

If participants reported a *specific* security-related behavior change, we asked them to explain further how the change was catalyzed—specifically, to discern between *social* and *non-social* catalysts for behavior change. Either way, we asked participants to explain, in detail, the context surrounding their decision to enact the change: Was the change brought about by a personal negative experience, or because of an article they read online? If they heard about a security incident through a friend, how did the friend broach the conversation? And, if a social process drove the change, we asked participants to clarify how the social process manifested—for example, did they seek out advice, or did a friend offer them unsolicited advice? We also asked participants whether and why they did or did not share their concerns, advice, or behavior change with anyone else.

We also asked participants if they could recall *specific* conversations they had about security and privacy. Did they ever share information about security or privacy? If so, what did they share, with whom, and why? By focusing on specific conversations about security and privacy (e.g., "I told my mother to update her privacy settings"), rather than general conversations (e.g., "People usually tell me to update my password"), we were often able to uncover the specific context of a conversation (e.g., a catalyst and goal for the conversation).

To capture security-related conversations that did not fit into the pre-constructed themes of mobile authentication, app installation, and social media privacy settings, at the end of the interview, we also asked participants more open-ended questions about conversations related to security and privacy.

We iteratively refined our protocol by piloting it with 5 people. All interviewers participated in the pilots in order to mitigate variation in delivery across interviewers and interview sessions. Questions that participants could not easily answer (e.g., hypotheticals) were culled through these iterations. Ultimately, our interview lasted approximately 45 minutes, and interviewees were compensated $10 to participate.

3.2 Recruitment

We recruited participants from an online recruitment tool that pairs research participants from the local area with research projects of interest. Participants were required to own a smartphone running Android or iOS, be an active user of any

social media service, and be at least 18 years old. We went through three rounds of recruitment to recruit a variety of occupations and ages across our sample. For example, in our first round of recruitment, we predominantly interviewed students in their mid-twenties. Thus, in subsequent recruitment rounds, we specifically recruited older non-students. We stopped recruiting additional participants once we believed we had sufficient diversity in occupation, age, and security proficiency to capture a large cross-section of experiences with security-related behavior change and communication. In our case, we appeared to reach this point after interviewing 19 participants—indeed, after the first 15, every additional participant echoed experiences very similar to those previously reported by others. Our recruitment solicitation is attached in Appendix B.

Our participants ranged in age from 20 to 54 years old (m=28.5, sd=10). Seven out of the 19 participants were female. Furthermore, as we tried to recruit participants from diverse backgrounds, 10 of our participants were non-students from many different professional backgrounds. All participants used an Android (n=12) or iOS (n=7) smartphone and were frequent Facebook users. Fifteen of the 19 participants reported using Facebook daily, while the remaining 4 reported that they checked Facebook at least a few times every week. Table 1 summarizes participant demographics. A more detailed description is in Table A1 of Appendix A.

3.3 Data Coding and Analysis

We recorded and transcribed, with consent, each interview, and used a qualitative data analysis program called Dedoose [37] to analyze the anonymized transcripts. We first partitioned each transcript into two sets of "excerpts". The first set of excerpts was a collection of all instances of an *action taken*, a *decision made*, or, more generally, a *behavior changed* related to security or privacy. As such, we will refer to this set of excerpts as the *behavior changes*. A representative example of behavior changes is P18's decision to rub-off the smudges on his Android device after a friend demonstrated that the smudges on his screen makes it easy for others to "crack" his Android 9-dot pattern:

"What I've been doing, I believe, after that scare with the nine dot, pretty much every time I turn off my phone, I put it in the pocket, I just kind of rub, just rub the smears off so you can't really see what direction I was going." (P18)

The second set of excerpts was a collection of all *specific* instances of communication about security and privacy, which we will refer to as the *communications*. An example excerpt comes from P14. After he received spam mail from a friend's e-mail account, he mentioned:

"I told my friend that this is something weird that came from your account. This is not what you would be probably into." (P14)

In total, from our 19 transcripts, we extracted n=114 behavior change excerpts, and m=118 communication excerpts. Excerpts were usually just answers to pointed questions, but to ensure robustness, two of the research group mutually agreed on all partition points for each excerpt.

We used these excerpts as our units of analysis—though, occasionally, we aggregated data across participants where it made sense (e.g., in determining how many participants actually changed their behavior as a result of a social process). We used an iterative, open coding process [21] to code the data, constructing codes where patterns naturally emerged and refining the codes iteratively until we reached consensus. Ultimately, we had two goals in mind through the coding process. The first was to understand the effect of social influence in driving *behavior changes*—which, in turn, means understanding the effect of social influence in modulating *security sensitivity*; and, the second was to better understand the triggers and reasons underlying *communications* about security and privacy.

Concretely, two researchers independently and openly coded a random subset of 20% of the excerpts from each of the *behavior changes* and *communications* excerpts. These openly generated codes were collaboratively synthesized into a set of high-level codes that three of the research team then used to code the remaining excerpts. Upon completion, the coding team discussed potential extensions to the coding scheme that arose from coding the new examples. If a change to the scheme was made, the coding team re-coded the full set of excerpts with the new scheme. We required two coding iterations to come to consensus.

From the 20% overlap of excerpts, overall inter-coder agreement was 85% for *behavior changes*, and 79% for *communications* (calculated as the number of overlapping excerpts where codes matched divided by the total number of overlapping excerpts). In cases of discrepancies, the coders discussed the discrepancies until agreement was reached, following standard practice. Inter-coder agreement for *each* applied code can be found in Table A3 in Appendix A, and all exceeded the 0.7 threshold commonly held to be acceptable in qualitative research [21].

4. RESULTS

4.1 Behavior Changes

First, we wanted to know if social processes often drove security related behavior changes, so we coded each *behavior change* excerpt as being driven by a *social* or *non-social* process. Excerpts were coded as being driven by a social process when the reason for the behavior change was social, and, importantly, if the social process was *clearly* reported by the participant in the transcript. For example, when asked about why he first enabled a PIN on his iPhone, P6 stated:

"When I first had a smartphone I didn't have a code, but then I started using one because everyone around me I guess had a code so I kind of felt a group pressure to also use a code." (P6)

As the underlying reason for the behavior change was a social process (observing one's friends) and was stated as such, we coded that behavior change as social. An example of a non-social behavior change comes, again, from P6. When asked why he changed his Twitter password, P6 responded:

"Diversification of passwords. I had the same password for every service so I wanted to pick a stronger password for… the service, yeah." (P6)

While P6 *could* have learned about the need for password diversification from friends, as he did not explicitly confirm this speculation, we coded the excerpt as non-social.

In all, out of the 114 behavior change excerpts, we coded a substantial 48 as being explicitly driven by some form of social influence. Furthermore, most participants (17 out of 19) reported at least one action taken, decision made, or behavior changed that was driven by social influence. Of note, however, is that the 48 examples of socially driven behavior change did not come uniformly from all of our participants. Notably P2 and P10 reported the largest number of socially driven changes at eight,

Trigger	N	Description	Example
Observed friends	14	Observing people around them engaging in a particular security behavior and emulated those people.	*"So when I was an undergrad I've been using it since then. And this four digit everybody started using it and it was a hype. And we had it." (P14)*
Social sensemaking	9	Discussing concerns with friends/loved ones to determine the right behavior.	*"I mean, like, one of my friends told me that you could alter the privacy settings so that, like, not everyone can look up your profile and not everyone can, like, try sending messages to you." (P15)*
Prank/ Demonstration	8	Friends/loved ones hacked into his/her account, demonstrating they were insecure.	*"Yeah, like my laptop was in my room. I walked out of my room and someone walked by and saw my Facebook and thought it would be funny to put something up." (P19)*
Security breach	6	Someone hacked into his/her account or information was shared too widely.	*"I did change that within the past week. The girlfriend was reading all of my mail, which is also a privacy concern" (P10)*
Sharing access	3	Sharing access to a device or account with another person leading to need for better security.	*"There are sometimes when you have to tell your friends what is my PIN number because they are a very good friend of yours and they have to make a call and I can't go every time and just unlock this for them." (P14)*

Table 2. Social triggers for behavior change derived from our iterative open coding process.

each. It is important to keep this bias in mind in any quantitative interpretation of our findings.

In all, these results suggest that social influence already plays a strong role in driving security and privacy related behavior change—even without any explicit social interventions. Next, we wanted to understand when and how social influence is effective at driving these behavior changes.

4.1.1 Social Triggers in Driving Behavior Change

To explore *when* social influence drove behavior change, we open coded the *triggers* for behavior change excerpts coded as "social". We found five primary social triggers for behavior change: *observing friends*, *social sensemaking*, *pranks and demonstrations*, *experiencing security breaches*, and *sharing access*. Table 2 lists all triggers, their frequency and their description. Next, to answer *how* social processes enacted behavior change, we also coded whether or not the socially driven behavior change examples in our dataset affected any part of the security sensitivity stack. Specifically, we asked the following:

Raised Awareness: Did the social process raise the participant's awareness of a new threat and/or security tool?

Raised Motivation: Did the social process raise the participant's motivation to protect him or herself against a security threat?

Raised Knowledge: Did the social process raise the participant's knowledge of how to use a security tool or method?

Importantly, we only answered "yes" to those questions if the *social* process mentioned in the excerpt was the reason for the heightened security sensitivity. For example, P16 mentioned that his Facebook account getting "hacked" resulted in him changing many of his passwords every 6 months at the advice of his friends, who he sought out for advice after the incident. In this example, the social process of P16 speaking with his friends raised his *knowledge* but not his *awareness* or *motivation*. It was the non-social process of experiencing a security breach that raised both his awareness and motivation.

For most (44 of 48) reported examples of socially driven behavior change, we found that the social process triggering the behavior change did, in fact, raise *some* form of security sensitivity. In fact, many examples raised *all* points of the security sensitivity stack. For example, P18 recalled advice he received on password composition after asking his friend to share a password:

"When I was working this summer, one of my co-workers told me about the whole algorithm thing. One, it just helps you I guess have different passwords. It helps you recall them easier based on I guess the type of profile. I guess you can cater, you can change your algorithm, depending on I guess what you want to be in it. But ever since I started using it." (P18)

In this example, the social process of P18 asking his friend about how to compose a password increased his *awareness* of a new method of password composition, his *motivation* to update his own method of password composition, and his *knowledge* of how to improve his method of password composition.

In the text to follow, we describe each social trigger we found in our data for security related behavior change. Furthermore, as a descriptive aid, we plotted how frequently different social triggers raised the different components of security sensitivity in Figure 1.

Observing friends (14/48 examples)

Most frequently, our participants reported changing their behavior after observing the actions of friends or others around them. In other words, participants changed their behavior after finding *social proof*—or, cues on how to act based on the actions of others [6]. For example, one participant in our sample adopted the 9-dot authentication method on his Android phone because his friends also used it. Additionally, as previously illustrated, P6 adopted a PIN because he felt "group pressure" to do so after observing everyone around him use authentication. This finding appears to be well supported by the background literature on technology adoption, which lists *observability* as a key criteria for an innovation to spread rapidly through social channels [25].

In certain cases, other forms of social influence apart from social proof appeared to be at play—specifically the social influence concepts of *liking*, or our tendency to follow the advice of those we like and those like us, and *authority*, or our tendency to follow the advice of those we consider to be authority figures [6]. For example, one participant indicated that she adopted a PIN code for her iPhone wholly because her mother, who she considered technically savvy, also had a PIN:

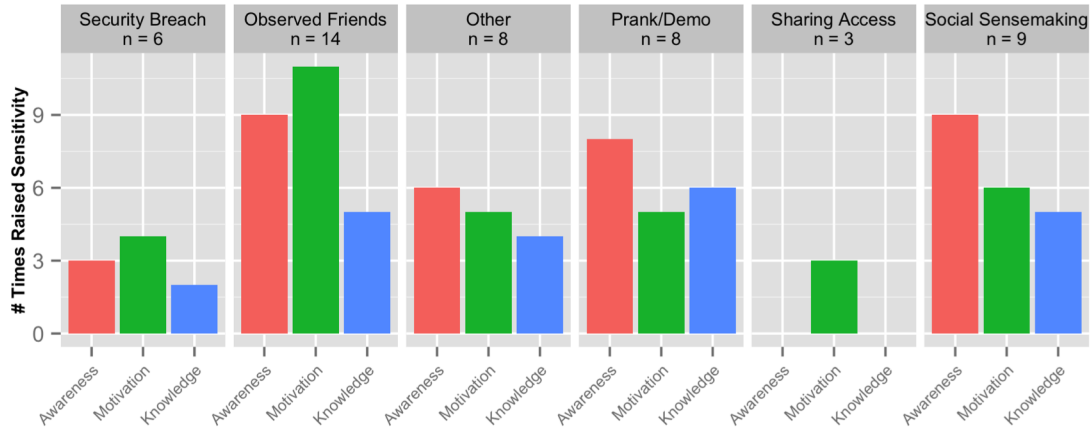


Figure 1. The number of times each social trigger for behavior change reported by our sample raised any of the three parts of the security sensitivity stack: awareness, motivation, or knowledge.

"My mother has-- she had an iPhone before I did, so she always had the block on hers, so I just kind of the... I think just because I saw her doing it, so it kind of just felt like it was something I had to do too." (P3)

Observation influenced behavior change for mobile authentication more often than the other specific topics we asked about in our interviews, probably because it is relatively easy to observe others authenticating onto their phones compared to observing others update their social media privacy settings or uninstall an app.

Looking at Figure 1, participants who observed others use security tools often were themselves *motivated* to start using those tools (11/14 examples). Furthermore, participants often became more *aware* of security tools after observing others' using those tools (9/14), but only occasionally gained *knowledge* of how to use the observed tools and methods by observing others (5/14).

Social Sensemaking (9/48 examples)

The second most frequent social trigger reported by our sample was *social sensemaking*—or, the process of making sense of a security system, tool, or threat by discussing concerns with others. We termed these triggers social sensemaking because they were similar in form and purpose to discussions, observed by Weick et al., among members of an organization who attempted to resolve uncertainty about recent novel events in their environment [33].

Participants often reported having discussions to resolve ambiguity in news and hearsay about security. The aim of these discussions was usually to find the correct or appropriate way to act to achieve the desired level of privacy or security within a system or with a security tool. In many cases, these discussions were prompted by a sudden infusion of uncertainty—for example, news articles about a novel security threat or gossip about anomalous security breaches others had experienced. Participants discussed these novel threats with others to share information about the threat, assess its veracity, and determine whether and how to change their behavior in response. For example, one participant in our dataset reported becoming more restrictive with posting to Facebook in response to a sudden, alarming, but unclear threat of all timeline posts becoming public:

"So yeah. I recently, like, a day or two, day before yesterday, I went through an ordeal. I don't know if it's fake or it's real, but somebody mentioned that all his private messages, they became public. Like, his messages with a friend. And it was like he had never thought of putting it on wall. And it suddenly opened his Facebook and everything was on his…I don't know if it's a real thing. And somebody mentioned in a comment that it happened with him as well, few days back." (P16)

P16's example is another illustration of *social proof* based social influence affecting an individual's security behavior: facing an ambiguous threat, P16 observed his friends for cues on how to act.

Social sensemaking also occurred when a participant wanted to understand a particular function within a system—for example, Facebook privacy settings. This need for specific information resulted in discussion and information sharing that exposed novel functionality or methods for protecting oneself against threats—often increasing participants' *knowledge* about the system (5/9 examples) and eventually leading to behavior change as a result. For example, one participant updated his privacy settings after a discussion that revealed novel system functionality:

"I mean, like, one of my friends told me that you could alter the privacy settings so that, like, not everyone can look up your profile and not everyone can, like, try sending messages to you. As in you can go to the privacy settings tab. And then, you could actually change it. Because I didn't know that you could do it, before. I mean, I just thought that it was default that everyone could look at your profile." (P15)

Social sensemaking also made participants more *aware* of available security tools (9/9), and the discussions would frequently *motivate* participants to act on their newly acquired knowledge (6/9).

Prank/Demonstration (8/48 examples)

The third most prevalent social trigger for the behavior changes reported by our participants was pranks and demonstrations—i.e., friends or loved ones cracking participant's accounts and devices as a prank, or to demonstrate that they were being insecure. Often, these pranks were explicit demonstrations to prove to the victim that their current security strategy or behavior was insecure. For example, one participant in our sample described a co-worker breaking into his phone to show the vulnerabilities of 9-dot authentication:

"One of my, when I was interning, engineering company, one of my friends and a fellow intern came to my desk, just unlocked my phone. I was surprised. I was like, "Hey, how'd you do it?" He put it against the sunlight and he saw I guess the smudges my finger left. He just followed the direction. Yeah, he had access to my phone." (P18)

Other prank examples reported were simply driven by opportunity—for example, a friend gaining unauthorized access to the participant's account because they left their Facebook account open on an unprotected device. Indeed, several of our participants were motivated to change their security behavior after their friends accessed their social media accounts and posted embarrassing information on their behalf. For example, one participant experienced this type of prank after leaving his laptop open and unprotected in his dorm room:

"Besides just my friends getting into my phone or on my Facebook and that's more from just me leaving my Facebook open or something if I walk out of the room and they just put up a funny status or something like or even just look through my messages or something like that. But nothing too threatening, more like practical joking side of it. But once that happens, I usually change my password immediately as would all of my other friends would too." (P19)

Pranks appeared to be quite effective at raising participants' security sensitivity. In all cases (8/8 examples), participants were made aware of a security threat and, in most cases, participants were instantly motivated (6/8) to update their behaviors to prevent a reoccurrence of the prank. Pranks aimed at demonstrating insecure behavior were also effective at raising participants' knowledge (5/8), as they were often followed up with direct or indirect lessons to prevent the breach from reoccurring—for example, the screen smudge "hack" reported by P18 taught him to wipe out the smudges from his phone screen periodically.

Experienced a security breach (6/48 examples)

Another prominent social trigger reported by our sample was experiencing a security breach—when participants or someone they knew had an account or device accessed by a stranger, or otherwise had information shared with unintended parties. In these examples, the victims of a security breach solicited advice from friends and loved ones, simultaneously spreading *awareness* (3/6 examples) of a new security threat, and *motivating* (4/6) behavior changes by grounding it in a real example of harm.

One participant initiated a new practice of updating his password on a monthly basis following his Facebook account getting breached, because his friend recommended that course of action:

"Because once I got my account hacked. And I was [doing my] bachelor's in a city, so yeah. After that I was more precautious regarding the same. And I'll keep changing my password, so on a monthly basis [because] My friends, actually they recommended me to do so. Like there's one of my friends used to do it. He said it's better to be safe than sorry, so…" (P16)

Sharing access (3/48 examples)

Another general social trigger reported by our participants was behavior change triggered by sharing a device or account with a friend or loved one—for example, modifying a password after allowing a friend to check their phone. These changes were a reflexive response to the fact that what participants desired to generally be private was now more widely available because of a transient need to share access. For example, one participant let her son use her phone and updated the passcode afterwards:

"One of my boys wanted to use my phone for something so I gave them my passcode. And not that I have anything that I don't care for them to see or anything, but after they did that then I changed it again because I just didn't want anybody to just-- I don't care if it's them or not. I don't want them to just be able to pick up my phone and do what they want with it." (P7)

While these triggers rarely raised awareness (0/3 examples) or knowledge (0/3), they seemed to be motivate participants to make a change (3/3).

Other triggers (8/48 examples)

Eight other instances of behavior change reported by our sample were triggered by other experiences, usually conversations or recommendations—for example, an authority figure recommending the use of authentication, as mentioned by P8 when asked why he first enabled mobile authentication:

"I think my boss at the time had it and he recommended it, because he leaves his phone at his desk." (P8)

Likewise, P10 mentioned adopting anti-virus software after receiving a recommendation from a friend who he considered a security expert, and P13 mentioned that she stopped using Google Chrome for financial transactions because two of her security expert friends informed her that the version of Chrome she used insecurely stored information. These recommendations often raised participants' awareness of, motivation to use *and* knowledge of how to use a new security tool or method.

Importantly, however, recommendations from authority figures didn't *always* result in behavior change. P13, for example, mentions that she ignored her boss's advice to have different passwords for different accounts because it would be hard to remember all those passwords. Nevertheless, the advice did raise her *awareness* of proper security practices.

P7 reported re-activating the PIN for her iPhone because a family member asked her why she deactivated it in the first place, urging her to reconsider. The conversation didn't raise her awareness or knowledge, but re-upped her motivation to use a security tool with a bit of social proof.

Interestingly, another participant mentioned installing anti-virus software on her laptop simply because she felt guilty, after conversing with others who attended her university's cybersecurity awareness fair, for not using software that her school provided:

"*I also felt guilty that I have all this free stuff I could install to protect my computer, and all this stuff I could do that's smart and I wasn't taking it." (P12)*

The guilt inspired behavior change reported by P12 is emblematic of the *reciprocity* principle of social influence, which suggests that people are more likely to follow the suggestions of those who did them a favor—even an unsolicited one [6].

Importantly, one participant reported how a social process urged her *against* behavior change (but was still responsible for a decision she made about security). P17 mentioned that she did not follow her security-expert husband's advice to delete unused and obscure online accounts because she noticed that her friends, who did not follow the advice, never experienced a security breach:

Catalyst	N	Description	Example
Observed Insecure Behavior	15	Noticed that someone was being insecure.	*"Right now I have ignored this storing passwords on my cell phone. He was like, 'Don't do this. It's dangerous.'" (P7)*
Observed Novel Behavior	11	Noticed a new security tool / method.	*"[I] see a lot of fancy password protection programs on [my co-workers] laptops. Like special files being encrypted. I'm like, "What's going on?" (P11)*
Sense of Obligation	15	Shared information out of obligation to protect others.	*"When I was younger, I remember my parents always telling me, like I'm sure everyone's parents tell them, to be very careful about who they give their Social Security number to. So, that's always like in my head, like if someone asks me for that, I'm just like, uh, no." (P14)*
Negative Experience	33	Experienced a security or privacy breach	*"Yes, my data got stolen. My photo got stolen on Facebook. I spoke to a couple of my friends. The only thing I could do was report abuse." (P6)*
Configuration	14	Had to set up security for a new device, account or security tool.	*"He was asking about Facebook, and he's a businessperson, so social media is somewhat of a new thing to him, and I think Facebook was-- he was just curious about it and how he could use it to kind of help his business and stuff like that. So…" (P20)*
News Article	15	Read a news article.	*"Well, before, I did not even know like I need to pay attention to this. Like I was aware of this, but I just did not know it was such a big deal. Then later, like I saw a topic, like online articles talking about that _____, talking about that, and that's when I went to the setting of like Facebook to change some." (P5)*

Table 3. Conversation catalysts derived from our iterative open coding process.

"I don't think it will be dangerous. Maybe I didn't see this kind of news or my friend didn't get some trouble when they didn't set password. Like, my friends sometimes they usually have a lot of different accounts, the same as me. But they didn't get any trouble. So I think maybe it will not be dangerous." (P17)

In this way, P17's friends' *lack* of a security breach offered her *social proof* that it's okay to ignore her husband's security advice.

4.1.2 Summary & Discussion

In summary, we analyzed 114 examples of behaviors changed, actions taken, or decisions made related to security and privacy, and found that social processes drove many (48) of those changes. We identified five common triggers for these socially driven changes, and found that these triggers were effective because they often raised participants' security sensitivity—usually awareness and motivation, but occasionally knowledge as well. These findings lend some support to the notion that social influence, especially in the form of *social proof*, *authority*, *liking*, and *reciprocity*, can be potent in raising security sensitivity—a result that bolsters the implications of prior work [14,23,29].

But, it remains unclear: is socially driven behavior change related to security and privacy as common as it could be? Socially driven change is the result of an interaction between two or more individuals—but those interactions are rare in the domain of security and privacy. Indeed, when asked why he didn't share his concerns about the U.S. government's pervasive surveillance (NSA PRISM) program, P11 stated: *"That's one thing I will never talk about."* Similarly, when asked about whether he has warned friends about a malicious smartphone application he uninstalled, P9 stated: *"Especially online. In person, it depends on the context. It does become a boring subject."*

The realization that conversations about security remain rare—and, thus, so too does the potential for socially driven behavior change related to security—begged the question: Under what circumstances do conversations about cybersecurity occur? To answer this question, we explored the 118 instances of *communication* about cybersecurity reported by our participants.

4.2 Communicating About Security

To understand the conditions under which conversations about security and privacy occur, we open coded excerpts about communication to surface triggering events for the interaction (*catalysts*) and the goal of the conversation (*conversation goal*).

4.2.1 Catalysts for Security Related Communication

We observed six primary catalysts for security related conversations in our dataset summarized in Table 3.

Insecure behavior

Some participants started a conversation about security in response to observing what they believed was insecure behavior, such as a friend or family member oversharing on social media:

"One of the reasons we talked about it is because I saw so many people post things on Facebook. A lot of times it's unnecessary things, you know, like just what they did today, "Oh, I had an amazing day," or, "I had a great dinner," and I was just talking to my husband, like why they-- I don't understand like they do that, like why they like to post things on Facebook to so-called to share." (P5)

Observing novel behavior

Relatedly, participants reported broaching conversations after observing novel security behavior or technology—for example, a new, visually appealing authentication technique. For example, one participant was stopped in a coffee shop and asked about the 9-dot authentication on his Android phone:

"We were just sitting in a coffee shop and I wanted to show somebody something and [they said], "My phone does not have that," and I was like, "I believe it probably does." (P10)

Sense of obligation

Obligations or responsibilities associated with a social role also prompted conversations about security. For example, parents lectured their children about security and privacy best practices (see example in Table 3 above), and managers informed their employees about how to manage company data because it was a

Goal	N	Description
Notify / warn	32	Notify or warn others of a potential security or privacy threat.
Prank/ Demonstrate	5	Demonstrate insecure behavior by hacking into a friend's account or device.
Share solutions	14	Share solutions, tools, and best practices (e.g., sharing how one composes his/her own password).
Vent	8	Seek social support / commiserate the experience.
Offer advice	19	Offer specific advice to others (e.g., update privacy settings, change password).
Seek advice	18	Ask for specific advice about security / privacy.
Storytelling	12	Topic was interesting/shocking/otherwise made for a good story.

Table 4. Conversation goals derived from our iterative open coding process.

part of their responsibilities. One participant described this type of interaction with his boss:

"When I was at work, I was given some sensitive documents, and I was told I couldn't send them over e-mail. I had to use a flash drive to move them over, encrypt them, then send them in e-mail." (P18)

This obligation included, in addition, a university's desire to protect its students. For example, one student talked about her university providing security solutions and advice in an annual security fair that she attended:

"They give us LoJack and all these different things you can get at the computer center. So we did talk about that. Like, locking up our computers and changing our passwords and stuff and being careful with the Wi-Fi." (P12)

Negative experiences

Negative experiences were the most common catalyst for security conversations reported by our sample. Indeed, many participants reported having conversations with friends and loved ones after experiencing a security breach. For example, one participant sought advice from friends after she received a friend request, on Facebook, from a fake profile using her own picture (see example in Table 3). Her friends recommended she report abuse in response to the attacks.

Configuration

Another frequent catalyst for discussion about security and privacy was configuring security and privacy settings on a new device, application or account. For example, one participant reported asking a friend for advice when a Facebook application asks for access to protected information:

"So there are many applications and Facebook would say that if you want to access them, there's a pop-up saying, "Allow," like, it will access all your information and stuff. So I asked him if I should go for it or not, and he tells me if it's worth going. Like, "Is it reliable or not?" (P16)

In general, participants frequently started conversations when setting or re-setting Facebook privacy settings (P13, P14, P16). In addition, many participants reported parents or older friends initiating conversations when they were setting up new computers or social media profiles for the first time (P4, P10, P15).

News articles

News articles or other press about security and privacy breaches also frequently triggered conversations. For example, one participant read and subsequently shared an article on social media about how over sharing could lead to identity theft and, more darkly, black market organ trading:

"I know there's like news talking about girls they are just so crazy about telling people on the social media where they are every minute, what they are doing every minute. So some criminals they actually use the information and just like kind of how do you say they found the girl according to her shared information online every minute. […] So I shared this article just to let my friends see just don't do it very often because I saw some of my friends on Facebook she did this really often like telling everybody what she was doing and what she had and where she was and like that." (P2)

4.2.2 Conversation Goals

We next analyzed our communication excepts for conversation goals to better understand what the conversation initiator wanted to achieve from the interaction—was it to warn others about potential threats, edify others about security tools or seek advice on how to configure security settings? During our open coding process, we identified seven distinct types of conversation goals, summarized in Table 4.

4.2.3 The Interaction of Catalysts and Goal

The interaction of conversation catalysts and goals provided enough context to answer the question: *under what circumstances do conversations about cybersecurity occur?*

To identify the most frequent conversations, we ran a cross tabulation of catalysts and conversation goal. For brevity, we focus here on the six most prevalent and interesting combinations, summarized in Table 6. These six combinations grouped into two broad categories of conversations, distinct in terms of their catalyst, focus and goal—*warnings* and *teachings*.

4.2.3.1 Warnings

Warnings were meant to raise awareness of a specific, immediate threat that had come to the attention of the conversation initiator. These warnings took three forms, varying in their catalysts, but resulted in a notification about a novel threat: cautionary tales, targeted warnings, and spreading the news.

Cautionary tales (10/118 examples)

The most common catalyst-goal combination reported by our participants was what we called cautionary tales—a conversation triggered by a negative experience on the part of the conversation initiator (or someone close to the initiator), with the goal of warning friends and loved ones about the threat. These conversations often involved sharing information about a recent security breach so that others could judge if their accounts or

Name	N	Catalyst	Content
Warning			
Cautionary tales	10	Negative experience	Notify / warn
Targeted warning	7	Insecure behavior	Notify / warn
Spreading the news	8	News article	Notify / warn
Teaching			
Lecturing	8	Sense of obligation	Offer advice
Configuration help	8	Configuration	Seek advice
Social learning	5	Novel behavior	Share solution

Table 5. The most frequent conversations about security and privacy, based on the catalyst and content.

information were in any danger. In several cases the conversation was a response to an out-of-character behavior on the part of a friend or family member. For example, after receiving odd requests for money from a friend via e-mail, one participant notified this friend that his email account was likely breached:

"Because, when I opened the e-mail, it said that they were, I think, they were in England and they didn't have enough money to come back to the States so can you send us some money, wire us some money, over, yeah. And if I'm not mistaken, I was probably the first to contact them that they were hacked. I'm like, 'This isn't right. Something strange'" (P11)

In another example, after his girlfriend illicitly accessed his e-mail account, one participant spoke to his friends to let them know that she may have read their conversations:

"It was just like, 'Hey, [my girlfriend's] been reading through our mail, like our conversations and stuff,' […] She probably read some of our conversations, not like she's going to get into your accounts." (P10)

Targeted warnings (7/118 examples)

Another common catalyst-goal combination we found was one where the conversation initiator issued a warning about potential security or privacy threats after observing others engaged in what they believed was risky behavior—what we call *targeted warnings*. For example, one participant described a friend warning her about the danger of not having a passcode:

"I was having a conversation with somebody and they were saying, 'Don't you have your passcode on there anymore?' And I said, 'No, it's a pain in the butt.' And they said, 'Well, it'd probably be a good idea if y- especially if you like leave it lay around on your desk or something like that. Or even if you're out in the evening and you have it on your purse, which most people now when they're out they have this thing right on the table where they are that somebody doesn't come by and grab it or whatever. That way they can do whatever they want with it.'" (P7)

Spreading the news (8/118 examples)

News articles about security breaches often resulted in conversations we refer to as *spreading the news*—conversations where the initiator attempted to warn friends and loved ones about a security threat outlined in a news article. These conversations sometimes included advice on how to change behavior to protect oneself from the new threat, but were usually just meant to raise awareness that a threat existed. For example, one participant talked about his contacts on Twitter discussing stories about Facebook privacy concerns without giving advice:

"Oh. Yes. People have said constantly on Twitter about how Facebook, it's not private anymore. Which is ironic, because neither is Twitter. So I've seen that, but no one has showed a article about being secure like with NSA and stuff." (P4)

As with other warnings, these conversations were often motivated by a desire to protect. For example, one participant described sharing a link to an article, through social media, about a credit card breach in order to warn her loved ones to be careful. Indeed, when asked why she shared one such news article, P2 said:

"To ask my beloved to actually pay attention to these things, to make sure they're okay. Their bank accounts are okay, if they actually do some shopping that day." (P2)

Conversations prompted by news articles also sometimes led to sharing best practices or details of privacy and security behaviors.

"We were just generally sitting around and somebody was like, 'Oh, this is an article about Facebook privacy stuff again. Let's look at it' 'Do you use this,' or 'I use that,' and 'Oh.' So really just comparing notes is the best way I can put it. Like we weren't overly scrutinizing each other's things. But like 'I found this to be effective.'" (P10)

4.2.3.2 Teachings

The other broad category of conversations we found was *teachings*. Teachings involved sharing security best practices or edifying others on how to protect themselves from security and privacy threats. In contrast to warnings, these conversations focused on sharing specific information about behaviors to enact in order to *solve* an immediate problem or *avoid* a future threat. Three conversations fell into this category: lecturing, configuration help, and social learning.

Lecturing (8/118 examples)

Conversations we referred to as *lecturing* involved advising others about security best practices, usually because the initiator felt a sense of obligation. Several of these conversations were between parents and children. Initially, parents offered children advice—for example, to not over share on Facebook. When children were older, however, they tended to be the ones lecturing their parents about privacy and security best practices. One participant described the litany of advice he gave to his parents about what to do and what not to do:

"I mean, I've spoken to my mom and dad about it. Like, I've told them, like, because I've told them to also use the same features that I do. Like having screen locks for phones and being more careful about passwords. And not logging into public computers and just leaving them without signing out." (P8)

Another type of lecturing was managers lecturing employees about security best practices to protect company data. For example, one participant described her boss asking her to regularly update her password:

"Actually, this was given to me by my manager, with whom I used to work. So he's the one who told me about this. He was like you

should change your password because it contains confidential information." (P13)

Another participant described his boss asking him to encrypt confidential files and transmit them physically on a USB flash drive rather than through email (P18).

Configuration help (8/118 examples)

Conversations about *configuration help* consisted of a conversation initiator soliciting advice on how to configure security and privacy settings for a new device or account. For example, one participant described helping his mother set up her new laptop with the appropriate security settings to keep her information safe (P19). Another participant described encouraging his mother to enable 9-dot authentication on her new Android phone to make sure no one else could access it:

"I mean, just the same reason that people shouldn't just look into her phone. Because, like, if it does not have a button, anyone can just, like, unlock and look at her messages and stuff." (P15)

Most frequently, configuration help conversations were about setting up the Facebook privacy settings (P1, P3, P4, P8, P19).

"If anything maybe my mom. I'm not sure directly security issues but she doesn't really know how to do Facebook that much so she'll ask me questions about it, in general, like how to post or, I guess, how to remove herself from something or certain things like that. So, I guess, I have given her advice in a way, just given her a few basic steps of set this as this just so you don't have--you're not completely open and public." (P19)

Social learning (5/118 examples)

In *social learning* conversations, conversation initiators observed novel security or privacy behaviors or tools—for example, a new way to compose passwords (P9, P10, P18) or a new type of authentication (P8)—that led to questions that allowed others to share information about the behavior. These conversations were opportunities for experts or early adopters to boast about their solutions for solving common security problems. For example, P18 asked a friend about sharing his Amazon account password, prompting the friend to share his password composition method:

"When I was working this summer, one of my co-workers told me about the whole algorithm thing. One, it just helps you I guess have different passwords. It helps you recall them easier based on I guess the type of profile. I guess you can cater, you can change your algorithm, depending on I guess what you want to be in it. But ever since I started using it." (P18)

4.2.4 Summary & Discussion

In analyzing the 118 conversations about security and privacy reported by our participants, we uncovered six common conversation catalysts (Table 3) and seven common conversation goals (Table 4). From these catalysts and goals, we identified six common catalyst-goal contexts (Table 5) that captured a large number of the security conversations reported by our sample, enabling us to answer the question: *under what circumstances do people generally talk about privacy and security?*

Broadly, the answer appears to be: to *warn* or to *teach*. Indeed, most commonly, our participants reported conversations about privacy and security to be educational experiences—either in sharing and receiving information about a novel security threat, or in sharing and receiving advice about how to solve a specific security problem or security best practices. This finding appears to confirm the notion that social processes *can* contribute to the heightening of security sensitivity, as these educational conversations often raised any or all of awareness, motivation or knowledge about security.

Observability, again, appeared to be a key driver of conversations—whether experts witnessing *insecure* behavior or non-experts witnessing *novel* behavior. In general, however, social learning may not have been as prevalent as would be ideal. Social learning conversations may represent the ideal context under which social influence *can* affect security sensitivity—novices interested in learning about security voluntarily ask for information from experts, thereby raising their own knowledge. In turn, experts are willing to share their information and don't feel that their efforts are wasted, as was implied by several of the security savvy participants we interviewed when asked why they don't share information about threats more often (P4, P9).

Unfortunately, many of our participants alluded to an illusory correlation [4] between security feature usage and paranoia, referring to their expert friends as "hyper-secure" (P5) and their actions as "above and beyond" (P18) or "nutty" (P1). Perhaps as a result of this negative perception towards those with high security sensitivity, many of the security savvy participants we interviewed mentioned that they avoided sharing information with their friends because the topic seemed socially inappropriate or unwelcome—as too preachy, for example. There is, thus, a substantial missed opportunity for experts to share knowledge with novices that only appears to be overcome when novices observe and query about interesting, novel behavior by the expert.

5. GENERAL DISCUSSION

Our results introduce a typology of social interaction around cybersecurity behavior and communication. First, we confirmed that social processes are an important influence on cybersecurity behavior change—a large number of behavior changes reported by our sample were driven at least partially through social processes. Specifically, we identified five common social triggers for security related behavior change—observing and learning from friends, social sensemaking (discussing ambiguous security threats with friends to determine the relevance of the threat and a clear course of action), pranks and demonstrations, experiencing a security breach and sharing access to a device with others. Furthermore, all social triggers for behavior change reported by our sample appeared to heighten security sensitivity in some way—either by increasing participants awareness of a new threat or security tool, motivating participants to protect themselves, or increasing participants knowledge of how to protect themselves.

We also found that conversations about security are primarily educational in nature, instigated mostly with a goal to *learn* or to *teach*. Many of our participants, for examples, reported having conversations about security to warn their friends and loved ones to be careful after experiencing a security breach, reading about a security threat on the news, or observing a friend's insecure behavior. Others reported specifically querying for security knowledge and advice after observing novel security behavior (e.g., the use of a new type of authentication), or if they had a specific and immediate security problem they wanted to solve (e.g., configuring the security settings of a new laptop).

Our results also emphasize the influential nature of a specific negative experience in raising the security sensitivity and, in turn, changing the cybersecurity behavior of victims and those around them. Interestingly, friends and loved ones appeared to at least indirectly take advantage of this fact, often breaking into others'

accounts to prove to that person that s/he was not fully protected. This notion of pranking by friends and family can also be considered as an effective way to create a *teachable moment*, something that past work on PhishGuru has found to be effective in teaching people about phishing attacks [20]. In other cases, pranks were not necessarily meant to directly educate victims, but were used as a form of "hazing". Either way, the breach elicited a similar reaction—both the victims of these negative experiences and the people around them who they shared the experience with became more aware of and motivated to address their own security vulnerabilities. These breaches also motivated participants to communicate with others to solve their problems.

The observability of security features and methods also proved to be important in driving behavior changes through social processes. Indeed, observing friends was the most frequent social trigger for behavior change. Nevertheless, most security features and methods are inherently unobservable and were rarely surfaced in our interviews—password composition methods, for example. When P18 learned of a new way to compose passwords from his expert friend, he immediately started utilizing this new composition policy. However, only two of our participants mentioned talking about password composition policies, suggesting there is much room for improvement in leveraging social processes to raise security sensitivity.

Observing novel or insecure behavior was also a key trigger for conversations about security and privacy, prompting novices to ask experts about novel behaviors and experts to warn novices about insecure behaviors. These conversations, again, were contingent upon the observability of the security feature or method. Experts could see the lack of mobile authentication on their friends' smartphones, but they could not see their friend's social media privacy settings, for example, and so conversations about social media privacy settings were rarely proactive—they were usually reactive, after someone encountered a breach.

However, simply increasing the observability of all security features may not be the best solution. First, security settings have historically been private—and for good reason. Indeed, past work by Gaw et al. [14] found that people who encrypted e-mail were often considered paranoid unless they were in a role where they handled sensitive company data, suggesting an illusory correlation [4] between security feature usage and paranoia. Our own interviews allude to a similar phenomenon, which appeared to be inhibit security experts from sharing their knowledge with others unless specifically asked. Indeed, as early adopters of security features are likely those who are especially concerned about their security—and, thus, are the most likely to be considered as paranoid by lay users—it is possible that making security decisions and behaviors perfectly observable might work *against* security sensitivity. After all, potential adopters may look at the present adopter list and find tenuous *social proof* that only "paranoid" people use a security feature. Second, we also saw evidence that social processes can work *against* a user following advice if it seems like none of their friends are affected by a threat. Likewise, it is possible that when a useful security feature has low current adoption, potential adopters might see the *absence* of adoption as *social proof* against using the feature.

To best leverage the positive effects of observability, therefore, it would seem that we want to facilitate more *social learning* conversations and *observing friends* behavior change. To that end, if we make security tools more visual and amenable to conversation while considering simple design for enhanced usability [36], non-experts can passively raise their *awareness* and *motivation* by observing their expert friends, and then raise their *knowledge* by voluntarily asking about security.

5.1 Limitations and Future Work

Our sample, although representative in many respects, is primarily from the US and young. Furthermore, as we solicited participants from only one online recruitment source, we could have introduced a systematic bias into our results—our participants were the type that generally volunteers for research projects. This means our results may not necessarily widely generalize, as is the case with most qualitative research. Thus, future work should examine whether the patterns and relationships identified in our data persist in a larger, representative sample of technology users. Our results are also limited to the communication and interaction instances participants could recall during our interview session—the so-called *recall* problem that afflicts retrospective interview studies [21]. Furthermore, as we only analyzed instances of behaviors changed, actions taken, and decisions made driven by social processes, we do not talk about the substantial number of non-social triggers for the same.

Our findings inform a breadth of potential future work, specifically in designing systems and interventions that *leverage* social influence processes to raise security sensitivity. For example, a key finding from our interviews was that the *observability* of security tool greatly facilitates its spread through social channels. Nevertheless, most security features are not observable, leaving little room for social spread and learning. Future work could introduce simple manipulations to increase the observability of security features and measure their effect on conversation frequency and behavior change, for example.

6. CONCLUSION

In summary, we qualitatively examined how social processes drive security-related behavior change and communications about security. Our findings suggest social processes played a major role in a large number of privacy and security related behavior changes reported by our interviewees, probably because these processes were effective at raising security sensitivity—the awareness of, motivation to use and knowledge of how to use security tools. In addition, conversations our participants had about security and privacy were most often instigated by the desire to (1) warn or protect others from immediate or novel threats observed or experienced and to (2) gather information about solving a privacy problem. One theme that arose from our interviews, especially, is that the *observability* of security feature usage was a key enabler of socially triggered behavior change and conversation—in encouraging the spread of positive behaviors, discouraging negative behaviors, and getting participants to talk about security. Altogether, our results suggest that there is a substantial and often overlooked social process that helps drive security related behavior change, and that in order to maximally raise security sensitivity, we should make security tool usage more observable and amenable to conversation. In addition, we believe our work provides a strong foundation for much needed further exploration into the social dimensions of cybersecurity behavior.

7. ACKNOWLEDGMENTS

This work was generously supported by NSF Award #1347186, the NDSEG Fellowship, and CMU's CyLab. We would also like to thank Samantha Finkelstein, Hsu-Chun Hsiao, and Ruogu Kang for helping with refining the interview protocol.

8. REFERENCES

[1] Acquisti, A. and Grossklags, J. Losses, Gains, and Hyperbolic Discounting: Privacy Attitudes and Privacy Behavior. In J. Camp and R. Lewis, eds., *The Economics of Information Security*. 2004, 179–186.

[2] Adams, A. and Sasse, M.A. Users are not the enemy. *CACM 42*, 12 (1999), 40–46.

[3] Bandura, A., Grusec, J.E., and Menlove, F.L. Vicarious Extinction of Avoidance Behavior. *Journal of Personality and Social Psychology 5*, 1 (1967), 16–23.

[4] Chapman, L.J. Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior 6*, 1 (1967), 151–155.

[5] Cialdini, R.B. and Goldstein, N.J. Social influence: compliance and conformity. *Annual Rev. of Psych. 55*, 1974 (2004), 591–621.

[6] Cialdini, R.B. *Influence*. Harper Collins, 2009.

[7] Davis, F.D. Perceived Usefulness , Perceived Ease Of Use , And User Accep. *MIS Quarterly 13*, 3 (1989), 319–340.

[8] Dhamija, R., Tygar, J.D., and Hearst, M. Why phishing works. *Proc. CHI '06*, ACM Press (2006), 581–590.

[9] DiGioia, P. and Dourish, P. Social navigation as a model for usable security. *Proc. SOUPS '05*, ACM Press (2005), 101–108.

[10] Dourish, P., Grinter, R.E., Delgado de la Flor, J., and Joseph, M. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing 8*, 6 (2004), 391–401.

[11] Egelman, S., Acquisti, A., Molnar, D., and Herley, C. Please Continue to Hold An empirical study on user tolerance of security delays. *Methodology*, (2010).

[12] Egelman, S., Cranor, L.F., and Hong, J. You've been warned. *Proc. CHI '08*, ACM Press (2008), 1065–1074.

[13] Furnell, S., Jusoh, A., and Katsabas, D. The challenges of undersatnding and using security: A survey of end-users. *Computers & Security 25*, 1 (2006), 27–35.

[14] Gaw, S., Felten, E.W., and Fernandez-Kelly, P. Secrecy, flagging, and paranoia. *Proc. CHI '06*, ACM Press (2006), 591–600.

[15] Goldstein, N.J., Cialdini, R.B., and Griskevicius, V. A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research 35*, 3 (2008), 472–482.

[16] Herley, C. and Oorschot, P. van. Passwords: If We're So Smart, Why Are We Still Using Them? *Financial Cryptography and Data Security*, (2009).

[17] Herley, C. So long, and no thanks for the externalities. *Proc. NSPW '09*, ACM Press (2009), 133–144.

[18] Inglesant, P.G. and Sasse, M.A. The true cost of unusable password policies. *Proc. CHI'10*, ACM Press (2010), 383–392.

[19] Kim, T.H.-J., Gupta, P., Han, J., Owusu, E., Hong, J., Perrig, A., and Gao D. OTO: Online Trust Oracle for User-Centric Trust Establishment. *Proc. CCS '12*, ACM Press (2012), 391–403.

[20] Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F., and Hong, J. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology 10*, 2 (2010), 1–31.

[21] Miles, M.B. and Huberman, M. *Qualitative Data Analysis: An Expanded Sourcebook*. Sage Publications, Inc., 1994.

[22] Milgram, S., Bickman, L., and Berkowitz, L. Note on the drawing power of crowds of different size. *JPSP 13*, 2 (1969), 79–82.

[23] Rader, E., Wash, R., and Brooks, B. Stories as informal lessons about security. *Proc. SOUPS '12*, ACM Press (2012).

[24] Renaud, K. Evaluating Authentication Mechanisms. In L.F. Cranor and S. Garfinkel, eds., *Security and Usability*. O'Reilly Media, 2005, 103–128.

[25] Rogers, E.M. *Diffusion of innovations*. New York, New York, USA, 2003.

[26] Sasse, M.A. Computer security: Anatomy of a Usability Disaster, and a Plan for Recovery. *Proc. CHI '03 Wkshp on HCI and Security Systems*, Citeseer (2003).

[27] Schultz, P.W., Nolan, J.M., Cialdini, R.B., Goldstein, N.J., and Griskevicius, V. The constructive, destructive, and reconstructive power of social norms. *Psychological science 18*, 5 (2007), 429–34.

[28] Sheng, S., Magnien, B., Kumaraguru, P., et al. Anti-Phishing Phil. *Proc. SOUPS '07*, ACM Press (2007), 88–99.

[29] Singh, S., Cabraal, A., Demosthenous, C., Astbrink, G., and Furlong, M. Password sharing. *Proc. CHI '07*, ACM Press (2007), 895–904.

[30] Stanton, J., Mastrangelo, P., Stam, K., and Jolton, J. Behavioral Information Security: Two End User Survey Studies of Motivation and Security Practices. *AMCIS*, August (2004), 2–8.

[31] Suo, X., Zhu, Y., and Owen, G.S. Graphical passwords: A survey. *Proc. ACSAC'05*, IEEE (2005).

[32] Wash, R. Folk models of home computer security. *Proc. SOUPS '10*, ACM Press (2010), 1.

[33] Weick, K.E., Sutcliffe, K.M., and Obstfeld, D. Organizing and the Process of Sensemaking. *Organization Science 16*, 4 (2005), 409–421.

[34] Whitten, A. and Tygar, J.D. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. *Proc. SSYM'99*, (1999), 14–28.

[35] Zhang, Y., Egelman, S., Cranor, L., and Hong, J. Phinding Phish : Evaluating Anti-Phishing Tools. *Proc. NDSS'07*, (2007).

[36] Zurko, M. E. IBM Lotus Notes/Domino: Embedding Security in Collaborative Applications. In L.F. Cranor and S. Garfinkel, eds., *Security and Usability*. O'Reilly Media, 2005, 607–622.

[37] Dedoose. http://www.dedoose.com.

Appendix A: Additional Figures and Tables

Expanded Demographics

	Age	Gender	Race	Occupation	Phone OS	Mobile Auth	Social Media Usage
P1	28	Male	African American	Customer Service	Android	None	Daily
P2	22	Female	Asian	Unemployed	iOS	None	Daily
P3	22	Female	African American	Student	iOS	PIN	Daily
P4	22	Male	African American	Student	Android	None	Daily
P5	27	Female	Asian	Unemployed	iOS	None	Daily
P6	29	Male	White	Software Developer	iOS	None	Daily
P7	54	Female	White	Administrative Assistant	iOS	PIN	Weekly
P8	31	Male	Indian	Unemployed	Android	None	Weekly
P9	30	Male	White	Software Developer	Android	None	Weekly
P10	37	Male	White	Graphic Designer	Android	9-dot	Daily
P11	54	Male	African American	Chef	Android	None	Weekly
P12	20	Female	African American	Student	iOS	None	Daily
P13	24	Female	Indian	Graduate Student	Android	None	Daily
P14	25	Male	Indian	Graduate Student	Android	PIN	Daily
P15	21	Male	Indian	Graduate Student	Android	9-dot	Daily
P16	22	Male	Indian	Graduate Student	Android	9-dot	Daily
P17	34	Female	Asian	Unemployed	iOS	None	Daily
P18	20	Male	African American	Student	Android	9-dot	Daily
P19	20	Male	White	Student	Android	9-dot	Daily

Table A1. Expanded participant demographics.

Co-Frequency of Catalysts and Reasons for Conversations

	Offer Advice	Share Solution	Vent	Seek advice	Notify or Warn	Storytelling	Prank or Demonstrate	Other	**Total**
Sense of Obligation	**8**	2	0	0	2	3	0	0	**15**
Insecure Behavior	4	0	1	0	**7**	0	2	1	**15**
Negative Experience	3	3	5	7	**10**	2	2	1	**33**
Configuration	2	2	1	**8**	0	0	0	1	**14**
News Article	1	0	0	0	**8**	3	0	3	**15**
Observed Novel Behavior	0	**5**	0	3	0	2	0	1	**11**
Other	1	2	1	0	5	2	1	3	**15**
Total	**19**	**14**	**8**	**18**	**32**	**12**	**5**	**10**	**118**

Table A2. Co-frequency of catalysts for conversations about security and privacy (rows) and reasons for starting the conversation (columns).

Inter-Coder Reliability for Each Applied Code

Code	Inter-Coder Agreement
Behavior Change: Social or Non-Social	0.93
Behavior Change: Trigger Event	0.87
Behavior Change: Raised Awareness	0.87
Behavior Change: Raised Motivation	0.80
Behavior Change: Raised Knowledge	0.80
Communication: Catalyst	0.71
Communication: Reason	0.86

Table A3. Inter-coder agreement of codes applied in our analysis, calculated from a 20% overlap of coded excerpts by two coders.

Appendix B: Recruitment Materials

We solicited study participants through CBDR, an online research study participation pool maintained by Carnegie Mellon's Department of Social and Decision Sciences. Below we show the posting for our study.

Study Name: ($) Talk to us about cybersecurity

Description:

Participate in an interview about how you learn about and manage online privacy and cybersecurity—for example, about mobile phones, passwords and social media privacy settings. We are researchers in the Human-Computer Interaction Institute at Carnegie Mellon University. We are studying how people learn about and manage cybersecurity. Please bring your smartphone and laptop for the study. We may ask you to show us your smartphone's home screen, and we may ask you to log into your Facebook account using your own laptop.

Eligibility: You must be (1) 18 or over, (2) a regular Android or iOS smartphone user, and (3) a Facebook user

Duration: 45 minutes

Pay: 10 Dollars

Privacy Concerns in Online Recommender Systems: Influences of Control and User Data Input

Bo Zhang
College of Communications
Pennsylvania State University &
Samsung Research America
buz114@psu.edu

Na Wang
College of Info Sciences & Technology
Pennsylvania State University &
Samsung Research America
nzw109@ist.psu.edu

Hongxia Jin
Samsung Research America
75 West Plumeria Dr.
San Jose, CA 95134
hongxia.jin@samsung.com

ABSTRACT
Recommender systems (e.g., Amazon.com) provide users with tailored products and services, which have the potential to induce user privacy concerns. Although system designers have been actively developing algorithms to introduce user control mechanisms, it remains unclear whether such control is effective in alleviating privacy concerns. It also is unclear how data type affects this relationship. To determine the psychological mechanisms of user privacy concerns in a recommender system, we conducted a scenario-based online experiment (N = 385). Users' privacy concerns were measured in relation to different data input (explicit vs. implicit) and control (present vs. absent) scenarios. Results show that a control mechanism can effectively reduce users' concerns over implicit user data input (i.e., purchase history) but not over explicit user data input (i.e., product ratings). We also demonstrate that control can influence privacy concerns via users' perceived value of disclosure. These findings question the effectiveness of user control mechanisms in recommender systems with explicit data input. Additionally, our item categorization provides a reference for future personalized recommendations and future analyses.

Categories and Subject Descriptors
Recommender

Keywords
Recommender system, Privacy, User Control, User Data Input, Privacy Concern, Information Disclosure

1. INTRODUCTION
Online recommender systems (e.g., Amazon.com, Yelp) have become unprecedentedly popular with the advancement of information tracking and prediction algorithms. Tracing extensive data about user preferences and behaviors, recommender systems can help users make better and faster choices specifically tailored for them in multiple areas of their lives (e.g., e-commerce purchasing, movie viewing, restaurant picking) [50]. This not only reduces users' cognitive load, but also provides them with more relevant and valuable services and products. Striving for more accurate predictions, a vast body of research has been devoted to creating and refining algorithms on recommender platforms [8, 30].

However, these personalized recommendations also pose severe threats to online users' privacy. To accurately predict what users want and need, recommender systems usually rely on a large amount of user data collected out of users' expectations [32], thereby inducing privacy concerns [3, 46]. The concerns, in return, affect users' evaluations of the system [29]. This user data includes demographic information that can point to one's unique identity (e.g., email addresses and social security numbers), as well as product-related footprints users leave online through web browsing and purchasing, hinting at one's tastes and habits. Due to the variation in sensitivity among numerous pieces of user data, it is inefficient to implement a holistic protection mechanism at the cost of recommendation quality. Hence, it becomes imperative to differentiate sensitive information from non-sensitive information and determine users' concerns about them in a recommender context respectively; that way, system developers can create suitable remedies for balancing prediction quality and privacy loss.

In addition to various types of user data, the channels they are collected through—either explicit (e.g., product rating) or implicit (e.g., purchase history)—also bring about privacy concerns [6]. Both approaches are meant to offer service providers extant data for predicting user needs. Explicit data input puts users in a conscious situation and requires their effort to complete the process, whereas implicit input is processed automatically, usually without user awareness. The former may empower users with a sense of control but make the privacy issue more salient, whereas the latter may provide users with more seamless convenience but also come with a sense of intrusiveness that leads to privacy concerns. This study examines how these two types of data input affect users' privacy concerns.

In addressing privacy concern issues in recommender systems, much attention has been put on creating solutions, such as granting users control over information release [31] or providing disclosure justifications [27]. In principle, control enables users to better manage their information flow and make decisions on information sharing, so as to reduce concerns about privacy. In reality, active user control could increase users' cognitive load, which may impede the expected effectiveness. Also, it is unclear whether the presence of a control mechanism will moderate the effect of data input on privacy concerns.

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9-11, 2014, Menlo Park, CA.

In addition to investigating the effects of data input and user control on privacy concerns, this study also probes the underlying psychological mechanisms that could explain the causes of privacy concerns in a recommender system. Such a mechanism has rarely been documented in prior work. Specifically, we focus on two constructs—*value of disclosure* and *trust*—as potential explanations for privacy concerns about different types of information in a recommender context.

Through exploratory factor analyses, the current study divided 21 pieces of demographic information into identifiable vs. unidentifiable information, and divided 26 types of products into sensitive vs. non-sensitive categories. Based on the preliminary categorization, we tested the effects of user control and data input in a popular recommender system (i.e., Amazon.com) via an online experiment with four different scenarios. We found a significant influence of control on reducing users' concerns about both types of information. For product-related information, if data is accessed by the recommender implicitly (i.e., through purchase history), the presence of user control plays a significant role in decreasing privacy concerns. However, if product data is collected by the recommender explicitly (i.e., through product rating), user control does not help to alleviate users' privacy concerns over information releasing. In addition, we found that value of disclosure, rather than trust, explains the underlying psychological mechanism of control's influence on privacy concerns.

This paper makes three main contributions to privacy research in recommender systems. First, we created two item-based information indices based on users' privacy concerns (i.e., one for demographics and one for product-related information). Our indices extended previous research [27, 47] by extracting new factors. Based on these categorizations, future system developers can strategically adopt data input methods and privacy protection solutions. Second, our findings showed that user control was effective in reducing privacy concerns for implicit data (i.e., purchase history), but not for explicit data (i.e., product rating), which casts doubt on the current trend of embedding control mechanisms unconditionally for privacy-concern reduction. Thus, the implementation of a control mechanism in recommender systems should also be designed strategically. Last, adding to existing research on privacy in recommender systems, we propose *perceived value of information disclosure* as a psychological mechanism that could explain the phenomenon. We then discuss practical implications and directions for future research.

2. THEORETICAL BACKGROUND AND HYPOTHESES

Substantial research has highlighted the issue of privacy concern in recommender systems [3, 11]. In this section, we first review prior work on personalization and the relationship between data input type and privacy concerns. We then discuss the effectiveness of affording user control in alleviating such concerns in recommenders. Last, we consider psychological mechanisms that might explain users' privacy concerns. Built on extant previous privacy research, we also propose our hypotheses, research questions, and conceptual model.

2.1 Personalization and Privacy Concerns

Personalization, or proactive tailoring of products and services based on individuals' preferences and needs [9], is at the heart of recommender systems' functionality and technology [8]. It is of great importance to online vendors because user information can help them predict demand, build customer loyalty, and increase cross-selling possibilities [39]. Personalization also has been found to be of significant value to users, by providing convenience and better service matching [9], saving time and effort, and promoting an optimal user experience [28]. However, users may be hesitant to savor the benefits brought by sophisticated personalization technology [13] because these benefits inherently come with the sacrifice of some privacy. For example, personalized convenience may rely on unsolicited data collection [33], or the fact that recommender systems share user data with third parties [6]. This phenomenon is known as the "privacy-personalization tradeoff" [3, 9]. Some studies suggest that users rationally calculate the net value gained from information disclosure accounting for privacy loss [15, 52], whereas others argue they superficially process personalization cues on an interface [45, 54].

Regardless of how personalization is interpreted by users, its effectiveness mainly depends on two factors: the recommender's ability to capture and analyze user data, and users' willingness to share data and use personalized services [9]. The former may refer to different ways of collecting data (i.e., explicit vs. implicit data input) and also different types of data (i.e., demographic vs. product-related information); the latter points to an individual characteristic, namely the extent to which one values information disclosure in return for personalized benefits [25, 52]. Both aspects are addressed in this study.

2.2 Data Input in Recommender Systems

The efficacy of a successful recommender system is achieved by extensively acquiring, storing, and processing user data. This data varies in sensitivity and is gathered through different approaches. This study investigates users' privacy concerns regarding individual information items (i.e., demographic vs. product) used by a recommender system, and the influence of data input (i.e., explicit vs. implicit) on users' privacy concerns.

2.2.1 Demographic vs. Product Information

Recommender systems collect and analyze static demographic information that can be linked to individual identities (e.g., email addresses, social security numbers) [27] and dynamic online behavioral data that can infer one's tastes and preferences (e.g., purchase history, product ratings) [47]. Although this information has been deemed significant in affecting user privacy [19, 40], variation among individual information items in triggering privacy concerns has not been explored. Considering the vast number of footprints users leave online every day, and the difficulty in balancing prediction accuracy and privacy protection, it is critical to identify different types of user data that vary in sensitivity so that system developers can strategically implement different protection mechanisms.

There are two broad types of online user data: static demographic data and dynamic behavioral data. Past research has labeled them as "demographics" and "context" [27],

corresponding to two recommendation strategies: content filtering and collaborative filtering [30]. Behavioral data (i.e., recording what a user browses, clicks, and purchases online) is always associated with specific products. In an online shopping scenario, 23 product items were identified to raise different levels of privacy concerns, leading to reluctance in purchasing them [47]. Based on these previous definitions and findings, this study conducts an item-based privacy concern rating and analysis to evaluate how information type is connected with privacy concerns. Hence, we propose the following research questions:

RQ1a: What types of demographic information used by recommender systems for personalized recommendation will trigger privacy concerns?

RQ1b: What types of product information used by recommender systems for personalized recommendation will trigger privacy concerns?

2.2.2 Explicit vs. Implicit Data Input

To provide personalized recommendations, recommender systems rely on two kinds of user data input: explicit and implicit [30]. Past research has labeled them in various ways, for example, pull vs. push [48], overt vs. covert [52], customization vs. personalization (i.e., agentic actions vs. tailoring) [44], to name a few. Explicit input is direct feedback from users that clearly expresses their preferences and tastes, such as product ratings and movie critiques [10, 19]. Implicit input, on the other hand, is information unconsciously left by users online, which is often clustered automatically by algorithms to identify user-item connections for future recommendations [20]. Implicit input includes browsing history, purchase history, clicking behaviors, and search patterns [30]. Although users do not explicitly express their opinions for implicit data input, their tendencies can often be speculated based on their behavioral patterns. Explicit input requires users to be willing to give out information consciously, whereas user effort is not necessary for implicit input [40]. Therefore, the main difference between these two approaches lies in the presence of user consciousness and initiative. The two types of approaches are often adopted simultaneously in recommender systems for better prediction accuracy and efficiency [30].

Previous research has examined the effectiveness and impact of these input types on user perceptions and behaviors from different perspectives. For example, explicit input, rather than implicit input, has been found to be preferable in location-based advertising because users perceived more control and benefits in it and would also be more likely to employ it [48]. Implicit input may appeal to online vendors because they do not need to lobby users to opt in and because it may stimulate impulsive purchasing [48]. To users, however, implicit input can be intrusive because it means their data is tracked without consent; this could diminish the perceived value of recommendations, and even trigger negative reactions such as avoidance [16] and privacy concerns [7, 48]. In this study, we consider product rating and purchase history as representations of explicit and implicit data input, respectively. Drawing on this previous work about users' negative perceptions of implicit input, we posit the following hypothesis:

H1: In a recommender system, implicit data input (i.e., purchase history) will trigger greater privacy concerns about product information than explicit data input (i.e., product rating).

2.3 Empowering Users with Control to Reduce Privacy Concerns

Researchers, system developers, and policy makers have been creating solutions at all levels to cope with rising privacy concerns in online recommender systems and minimize the compromise of prediction accuracy. For example, Heitmann et al. [22] proposed an architecture that enables users to decide what personal information can be accessed by which service providers; Arlein et al. [2] designed a data protection mechanism that allows users to hide their real identities and use personae for information sharing; Xu et al. [53] demonstrated that privacy assurance approaches such as the TRUSTe seal can also reduce users' privacy concerns by way of perceived control over personal information. Most of these solutions endow users with either the capability of actually controlling their information sharing or with a perceived sense of control, which has been found effective in alleviating privacy concerns.

Indeed, the idea of privacy is often associated with control over personal information; if something is considered to be private, we want to be able to protect it [12, 43]. Many researchers directly define privacy as a sense of control [35, 49]. The notion of control is frequently studied as a key factor of privacy concern [34]. Loss of control over collection and usage of information has been found to lead to a greater sense of privacy invasion among online consumers [14, 41]. Milne and Boza [37] showed that, in general, individuals have less privacy concerns when they have a greater sense that they controlled the disclosure and subsequent use of their personal information. Acquisti and Gross [1] found that Facebook users who were not concerned about privacy of the information they posted online also felt a greater sense of control over it. Given the negative relationship between control and privacy concerns suggested by prior research, we propose the following hypotheses:

H2a: The presence of user control will lead to a decreased level of privacy concern about demographic information compared to the no-control condition.

H2b: The presence of user control will lead to a decreased level of privacy concern about product information compared to the no-control condition.

Because we also consider the effect of data-input type (i.e., explicit and implicit), which only applies to product-related information, we further ask the following research question:

RQ2: Is there an interaction effect between data input type and user control on privacy concerns toward product information?

2.4 Psychological Mechanisms of Privacy Concerns

Although the work discussed above provides insightful findings, it remains unclear which particular psychological mechanisms determine privacy concerns in a recommender context. Given numerous technological attempts in affording user control to reduce privacy concerns, we employ *perceived value of information disclosure* and *trust toward the recommender system* as potential underlying psychological paths.

2.4.1 Value of Disclosure to Users

The privacy calculus model posits that perceived value in information disclosure is often evaluated by weighing benefits and risks [14]. In a privacy context in recommender systems, then, perceived value of information disclosure can be defined as the trade-off between what users can gain from using the recommender and what risks users need to take in disclosing their information [52]. The deployment of a user control mechanism in a recommender system is likely to increase users' perceived benefits, as well as reduce their concerns about privacy loss. This leads to the following hypothesis:

H3: The presence of user control will lead to greater perceived value of information disclosure compared to the lack of control.

We define perceived value of information disclosure as the trade-off between benefits received from the recommender system and privacy loss to the system; therefore, the greater the perceived value, the more perceived benefits outweigh privacy risks, and the less likely one is to be concerned about privacy. As such, we further posit the following hypothesis:

H4: Perceived value of information disclosure mediates the relationship between user control and privacy concerns toward specific information.

2.4.2 A Trust-building Mechanism

Trust also has been found to be a concept that is closely associated with privacy in an online environment [18, 42]. Belanger et al. [5] found that consumers heavily rely on the trustworthiness of an online vendor to disclose their information; Milne and Boza [37] demonstrated that trust-building is more effective than concern-reducing in managing online users' information. Studies also showed that, at an institutional level, trust can significantly mediate the effect of privacy assurance practices on users' privacy concerns [14, 51]. Based on prior research, privacy coping strategies (e.g., providing user control) may assure users that their information will only be accessed and used with their consent, thereby inducing perceived trust toward the service provider, and eventually leading to a reduced level of privacy concern. Thus, we hypothesize the following:

H5: The presence of user control will lead to greater perceived trust toward the recommender system compared to the lack of control.

H6: Perceived trust toward the recommender system mediates the relationship between user control and privacy concerns.

2.4.3 Influences of Personal Traits

Personal traits, or individual characteristics, reflect human natures and can determine one's perceptions and behavioral patterns in many situations [38]. This is especially true in a privacy context because individual dispositions are often linked with one's privacy concerns and tendency to disclose information [4]. Three personal traits are particularly of relevance to this study: general privacy concern, perceived value of personalization, and perceived importance of control. To account for their potential influences on users' privacy concerns, these traits are all included as control variables.

2.5 A Conceptual Model

Grounded in theoretical research and prior empirical studies, we propose a conceptual model for the current study (Figure 1).

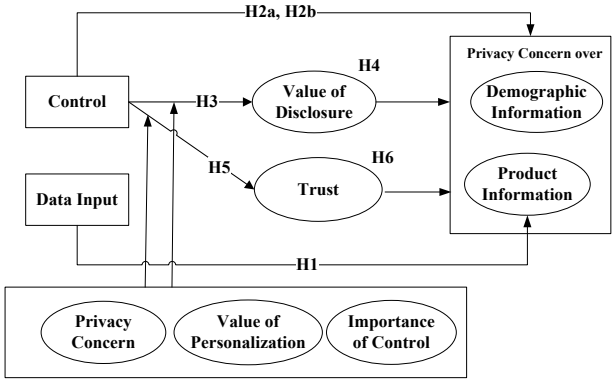


Figure 1. A conceptual model

3. METHODOLOGY

3.1 Study Design

The study's design consists of two components: item-based privacy concern ratings and scenario-based privacy concern probing. On the whole, a 2 (data input: explicit vs. implicit) x 2 (user control: presence vs. absence) scenario-based online user study was conducted to answer our research questions and test our hypotheses and conceptual model. To provide an index of information items that vary in sensitivity and differentiate them by degree of user privacy concern, we included 21 pieces of demographic information inspired by Knijnenburg et al. [27] and 26 product types [47]. Based on the scenario, participants were asked to rate how concerned they would be if Amazon.com accessed different types information for personalized recommendations. The purpose of this rating was to provide a relative measure of privacy concern on an item-by-item basis, rather than an absolute scale of information sensitivity.

3.2 Participants and Recruitment

We recruited all participants (N = 385) through Amazon Mechanical Turk (MTurk, www.mturk.com), a recruitment source that has become popular for conducting online user studies in recent years [26]. We restricted participants to US residents with a North American IP address and a Human Intelligence Task (HIT) approval rate of 90% or higher. Participants were also required to have made at least one purchase on Amazon.com in the past year so that the scenario setting would seem applicable to them. As an incentive, we paid each eligible participant 50 cents for a completed task. The majority of the participants were male (63.2 %) and Caucasian (75.8 %). The average age was 31.14 (SD = 10.70). We recognize the potential confounding effect of using an Amazon-based participant pool for an Amazon-related study. However, any confounding effects will be identical across conditions, so this should not cause any analytic problems.

3.3 Scenarios

We created four scenarios to examine the effects of user control (presence vs. absence) and data input (explicit vs. implicit) on

users' privacy concerns in a recommender system. All four scenarios were grounded in an online shopping context with Amazon.com due to its prominent role in recommender systems. Specifically, we instructed participants to imagine that they had purchased all of the listed products from Amazon.com. In the scenarios with presence of user control, participants were explicitly told that they had control over the extent to which Amazon.com could access their demographic information and purchase history in exchange for personalized recommendations; in the scenarios without user control, such information was not provided. Within these scenarios, we also varied two types of product-related data input by asking participants to evaluate their level of concern over releasing product information in their purchase history (i.e., implicit data input) or product ratings (i.e., explicit data input) to Amazon.com for personalized recommendations. Participants were randomly assigned to one of the four conditions/scenarios.

3.4 Procedure

After participants were pre-screened for eligibility, they were randomly assigned to one of the two user control scenarios (i.e., presence vs. absence), where they were instructed to evaluate their levels of concern over releasing 21 types of demographic information to Amazon.com in exchange for personalized recommendations. After that, participants were randomly assigned to one of the two data input scenarios (i.e., purchase history vs. rating), where they were asked to assess their privacy concerns over releasing 26 types of products purchased from Amazon.com in exchange for personalized recommendations. After the item-based privacy concern evaluations, we measured perceptual variables regarding users' attitudes toward the recommender system and individual characteristics. We then collected demographic information for use as control variables.

3.5 Measurements

To the extent possible, we adopted measurement scales for the main constructs in this study from prior research to fit the Amazon.com recommender context.

Inspired by Knijnenburg et al. [27], we included 21 pieces of demographic information that vary in sensitivity (e.g., email addresses, phone numbers, social security numbers). For product items, we included 26 products that also differ in sensitivity (e.g., textbook, hunting knife, bulletproof jacket) [47]. Privacy concerns about these items were measured on a Likert-type scale from "1 = not concerned at all" to "7 = extremely concerned."

Value of information disclosure was assessed by 3 items adapted from Kim et al. [25] and Xu et al. [52] (e.g., "The value I gain from use of Amazon.com's service is worth the information I give away.") (α = .756). Trust toward Amazon.com was measured with 6 items (e.g., trustworthy) (α = .886) [36].

In terms of individual differences, we measured participants' general privacy concern, perceived value of online personalization, and perceived importance of control. General privacy concern was measured via 3 items (e.g., "I am sensitive about giving out information regarding my preferences") [9] (α = .828). Value of online personalization was assessed with 6 items (e.g., "I value websites that are personalized for my usage experience preferences.") derived from Chellappa and Sin [9] (α = .855). Participants were also asked to indicate their perceived importance of control in the recommender context (e.g., "It is important for me to control the amount of information accessed by Amazon.com for personalized recommendations") (α = .943). This final measure was specific to the study and, therefore, created by the researchers.

All these measures took the form of 7-point Likert-type scales, with 1 being the lowest level and 7 the highest. A complete list of measurement items can be found in Appendix C.

4. RESULTS

We present our results by first describing the item-based analyses of privacy concerns over demographic and product information in response to our research questions. We then examine effects of data input and user control on participants' psychological perceptions and privacy concerns. Finally, we test our conceptual model of the psychological mechanism of privacy concerns in a recommender system via mediation analysis and structural equation modeling

To rule out confounding issues, we statistically control for general privacy concern, perceived value of online personalization, and perceived importance of control, along with other demographic information (e.g., gender, age, education).

4.1 Item-based Analyses of Privacy Concern

To discriminate sensitive information items from non-sensitive items and create an index of data types based on users' privacy concerns, we performed exploratory factor analyses (EFA). Table 1 shows a complete list of the 21 pieces of demographic information we included in the study, and Table 2 shows the 26 specific product types.

4.1.1 Demographic Information Type: Unidentifiable vs. Identifiable

The 21 items regarding privacy concerns over demographic information were first subjected to a principal axis factoring analysis (PAF) with an oblique, promax rotation. PAF was chosen because it generally produces outcomes close to maximum likelihood extraction and it is not overly sensitive to nonnormality [17]. An examination of the Kaiser-Meyer Olkin (KMO) measure of sampling adequacy suggested that the sample was factorable (KMO = .951). Scree-plot analysis indicated two factors for demographic information. The rotated pattern matrix of the item pool is shown in Table 1. One severely cross-loading item, *date of birth*, was dropped from the analysis based on the 0.3 rule (i.e., an item's highest loading should be at least 0.3 higher than its other loadings).

The 12 types of personal information that loaded onto Factor 1 represent general personal attributes that cannot be used as identifiers of a particular person. Hence, we labeled Factor 1 as "unidentifiable demographic information." On the contrary, the 8 items that loaded onto Factor 2 represent unique information that can be used to identify or locate an individual. Therefore, this was labeled as "identifiable demographic information." The individual items for each factor, factor loadings, and reliabilities can be found in Table 1.

To address RQ1a and test the difference in causing privacy concerns between the two demographic information types, a paired samples *t*-test was conducted. Results showed that users

were significantly more concerned about releasing identifiable demographic information (M = 3.850, SD = 1.571) to Amazon.com in exchange for personalized recommendations than unidentifiable demographic information (M = 3.188, SD = 1.469, $p < .001$).

Table 1. Exploratory factor analysis and privacy concern levels for demographic information

Component	Privacy Concern (Range: 1 to 7)	Factor Loading 1	Factor Loading 2
Unidentifiable Demographic Information (α = .938)			
Education	2.95	**.960**	-.170
Relationship	3.22	**.913**	-.177
Race	2.68	**.859**	-.097
Field of work	3.06	**.846**	-.005
Tech use	3.15	**.742**	.040
Interest	2.70	**.731**	.031
Gender	2.40	**.710**	.089
Age	2.65	**.688**	.174
Company	3.25	**.638**	.173
Calendar	3.66	**.604**	.040
Income	3.93	**.552**	.242
Web browsing	4.60	**.486**	.125
Identifiable Demographic Information (α = .904)			
Credit card	4.73	-.270	**.920**
Home address	3.57	-.088	**.919**
Phone number	3.88	.017	**.830**
Email	3.05	.051	**.756**
Name	2.94	.169	**.631**
Location	3.44	.232	**.616**
IP address	3.86	.113	**.585**
SSN	5.79	.022	**.463**
Dropped Item			
Date of birth	3.38	.384	.466

4.1.2 Product Type: Non-sensitive vs. Sensitive

In a similar manner, the 26 specific products tested were subjected to a PAF with a promax rotation. KMO suggested that the sampling was adequate for factor analysis (KMO = .967). Scree-plot analysis indicated two distinct factors for product types. The rotated pattern matrix is in Table 2. *Cigarette*, *lingerie*, and *bulletproof jacket* were dropped because of cross-loading, and *shoes* was discarded because of multicollinearity.

The 12 types of products that loaded onto Factor 1 are all products that are normally not considered to be sensitive, such as *office supplies* and *everyday necessities*. This factor was labeled as "non-sensitive products." The 13 types of products that loaded onto the second factor are products related to personal values and mental states, such as *HIV tests*, *depression-related books*, *bomb-making books*. Thus, we labeled Factor 2 as "sensitive products." Individual items for each factor, factor loadings and reliabilities can be found in Table 2.

To address RQ1b and examine how sensitive products are different from non-sensitive products in triggering privacy concerns, we conducted a paired samples t-test. Results showed that users' were significantly more concerned about releasing information about sensitive products (M = 3.762, SD = 1.768) to Amazon.com in exchange for personalized recommendations than non-sensitive products (M = 2.085, SD = 1.350, $p < .001$).

Table 2. Exploratory factor analysis and privacy concern levels for product information

Component	Privacy Concern (Range: 1 to 7)	Factor Loading 1	Factor Loading 2
Non-sensitive Products (α = .965)			
Furniture	1.83	**.972**	-.126
Food	1.85	**.969**	-.108
Flower	1.83	**.959**	-.099
Laptop	2.00	**.939**	-.064
Textbook	1.87	**.926**	-.055
Game	1.92	**.901**	-.046
Jewelry	2.07	**.867**	-.003
Peroxide	2.15	**.767**	.108
Hunting knife	2.35	**.678**	.229
Fertilizer	2.26	**.657**	.143
Weight loss product	2.48	**.614**	.303
Sensitive Products (α = .956)			
STD medication	4.61	-.248	**.978**
HIV test	4.58	-.173	**.942**
Sex toy	4.26	-.130	**.920**
Porn DVD	4.29	-.107	**.902**
Adult diaper	3.80	-.027	**.822**
Lubricant	3.37	.126	**.769**
Book-Bomb making	4.59	-.114	**.756**
Pregnancy Test	3.48	.139	**.732**
Book-Depression	3.23	.164	**.714**
Condom	3.26	.226	**.679**
Book-Bankruptcy	3.28	.220	**.655**
Dropped Items			
Shoes	1.83	1.011	-.175
Cigarette	2.66	.501	.349
Lingerie	2.95	.366	.526
Bulletproof Jacket	3.22	.309	.511

4.2 Effects of Data Input and User Control

Based on the level of privacy concern, users' demographic information can be classified into two categories: unidentifiable demographic information and identifiable demographic information. Similarly, product types can be classified into two categories: non-sensitive products and sensitive products. We adopt these classification results in the following analyses.

4.2.1 Effects of Data Input

In order to test the effects of user data input (explicit/rating vs. implicit/purchase history) in the recommender system on users' perceived privacy concerns, a series of analyses of covariance (ANCOVAs) were conducted, controlling for the influences of general privacy concern, perceived value of personalization, perceived importance of control, and demographics (e.g., age, gender, education). It is worth noting that data input type in recommender systems only applies to product-related information, not demographic information; product information can be obtained through user rating or history checking, whereas demographic information can only be obtained through user

input. Results showed a significant main effect of data input on value of information disclosure about product-related items, F (1, 376) = 7.85, p < .01. Specifically, participants perceived more value in disclosing purchasing history (M = 4.528, SD = .078) than in disclosing product ratings (M = 4.234, SD = .077) in exchange for personalized recommendations. However, data input's effect on privacy concerns about product information was not significant. Thus, H1 was not supported.

4.2.2 Effects of User Control

A multivariate analysis of covariance (MANCOVA) was conducted to investigate the effects of user control in the recommender system. Results indicated a significant overall main effect of user control, Wilks' Λ =.941, F (1, 372) = 5.839, p < .001. Subsequent univariate analyses showed that participants tended to express a higher level of perceived value of information disclosure [(M = 4.519, SE = .077), F (1, 376) = 6.629, p = .01] and significantly less concern about their unidentifiable demographic information [(M = 2.934, SE = 1.317), F (1, 376) = 7.863, p = .005], identifiable information [(M = 2.934, SE = 1.417), F (1, 376) = 22.345, p < .001], non-sensitive products [(M = 1.914, SE = 1.173), F (1, 376) = 4.752, p = .030], and sensitive products [(M = 3.492, SE = 1.714), F (1, 376) = 6. 352, p = .013], when they had control over information access then not (M = 4.236, SE = .078; M = 3.369, SE = .090; M = 4.178, SE = .097; M = 2.230, SE = .093; M = 3.974, SE = .118, respectively) (see Figure 2 (a) & (b)). However, the effect of control on perceived trust toward the recommender was not significant (F (1, 376) = .038, p = .846). Therefore, H2a, H2b, and H3 were all supported, but H5 was not.

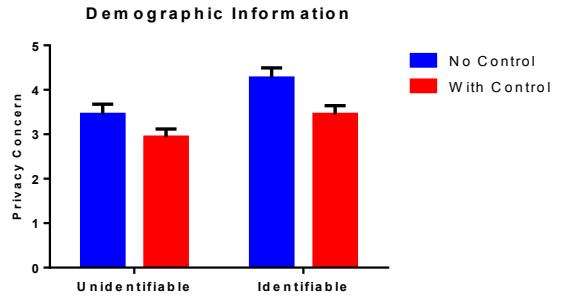


Figure 2(a). Effects of user control on privacy concerns about demographic information

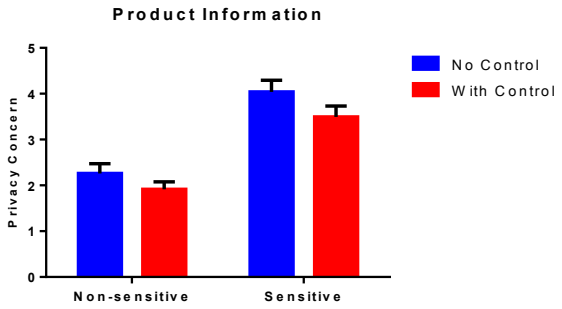


Figure 2(b). Effects of user control on privacy concerns about different information types

4.2.3 Interaction Effects between Data Input and User Control

To answer RQ2, we tested the interaction effects between data input and user control. We found that data input type significantly moderated the relationship between the existence of user control and perceived privacy concern about non-sensitive product information, F (1, 374) = 4.657, p = .032, but not sensitive product information, F (1, 374) = 1.691, p = .154. Specifically, empowering users with control over information release significantly lowered their concerns over purchase history containing non-sensitive products (M = 2.402, M = 1.914, for no-control and control conditions respectively). However, if users were asked to explicitly rate the non-sensitive products they had purchased before, such a control would not make a difference (M = 2.125, M = 2.133, for no-control and control conditions respectively) (Figure 3). This significant interaction effect did not exist for sensitive products.

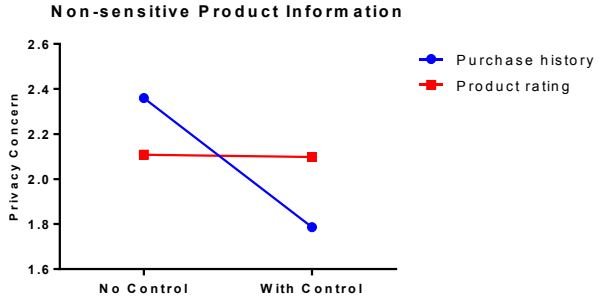


Figure 3. Interaction effect of user control and data input on privacy concerns about non-sensitive products

4.3 Testing the Conceptual Model

Before testing the overall conceptual model, we first speculated the degree to which effects of user control on privacy concerns over the four types of information (i.e., unidentifiable demographics, identifiable demographics, non-sensitive products, sensitive products) might be mediated by the two proposed psychological mechanisms—perceived value of disclosure and trust. An SPSS script developed by Hayes [21] was adopted to probe such mediation effects. As shown in Figure 4, perceived value of disclosure significantly mediated the effects of control on privacy concerns about unidentifiable demographic information (β = -.14, p < .001, Figure 4a), identifiable demographic information (β = -.12, p < .001, Figure 4b), non-sensitive products (β = -.08, p < .001, Figure 4c), and sensitive products (β = -.16, p < .001, Figure 4d). However, perceived trust toward the recommender system was not a significant mediator in any of these relationships. All path coefficients are shown in Figure 4. These findings provide support for H4 but not for H6.

Because we did not find any significant effect of user control on trust toward the recommender, nor did we yield a significant indirect effect of user control on privacy concerns via trust, we removed the trust construct from our conceptual model for the following statistical testing.

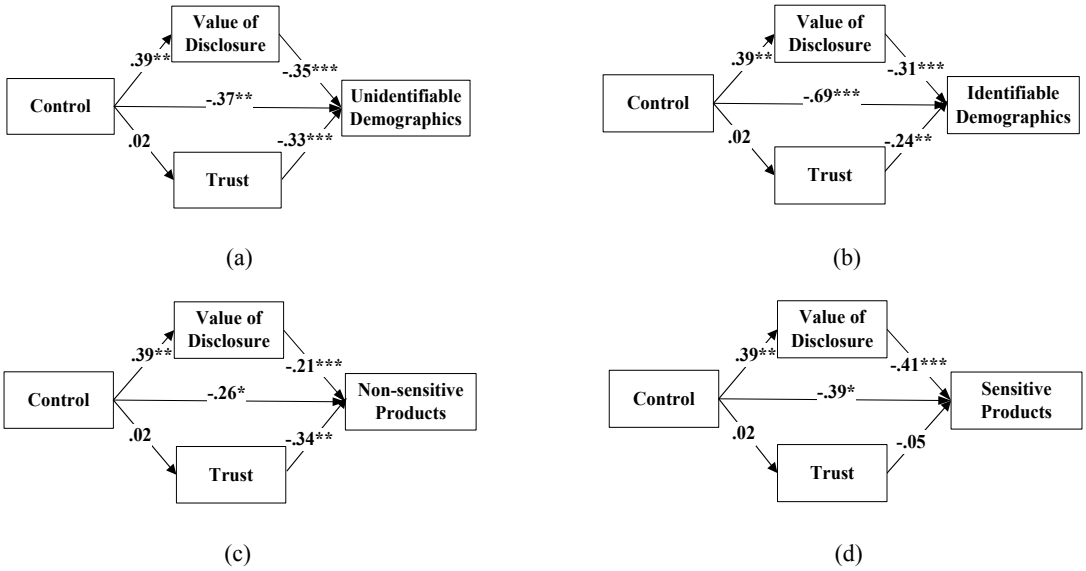


Figure 4. Path models for control's effect on privacy concerns about information [four types, (a)-(d)] with value of disclosure and trust as possible mediators (*p< .05, **p< .01, ***p< .001)

Given that user control plays an important role in affecting privacy concerns through perceived value of information disclosure, we tested the overall conceptual model with structural equation modeling (SEM) to map out the relationships among our main constructs. The 8-latent-factor structure with 57 individual items was found to retain a reasonably good fit: χ^2 = 9293.596, df = 3884, p < .001, root mean square error of approximation (RMSEA) = .049, 90% confidence intervals (CI): .048-.050, comparative fit index (CFI) = .825. And a subsequent multigroup structural equation modeling (MGSEM) with data input type (explicit vs. implicit) as the grouping variable yielded close good-fitting models. Figure 5 presents the final overall model and standardized path coefficients.

Consistent with previous findings, empowering users with control in the recommender system tends to enhance participants' perceived value of information disclosure. Such increased value of disclosure directly alleviates users' concerns over releasing their demographic and product-related information in exchange for personalized recommendations.

To further probe this effect, bootstrapping procedures were employed using 2000 bootstrap samples and a bias-corrected confidence interval in a multigroup analysis. With data input type as the grouping variable, results showed that the significant mediating effects of perceived value of disclosure only exist when data input is implicit (i.e., when purchase history is accessed for personalized recommendations). Specifically, perceived value of disclosure significantly mediated the relationship between the presence of user control in the recommender and privacy concerns about non-sensitive products (β = -.13, p = .006) and sensitive products (β = -.12, p = .009) when users thought their purchase history would be accessed. However, in the product-rating scenario, such mediating effects were not significant (β = -.06, p = .16; β = -.06, p = .12; for non-sensitive products and sensitive products respectively).

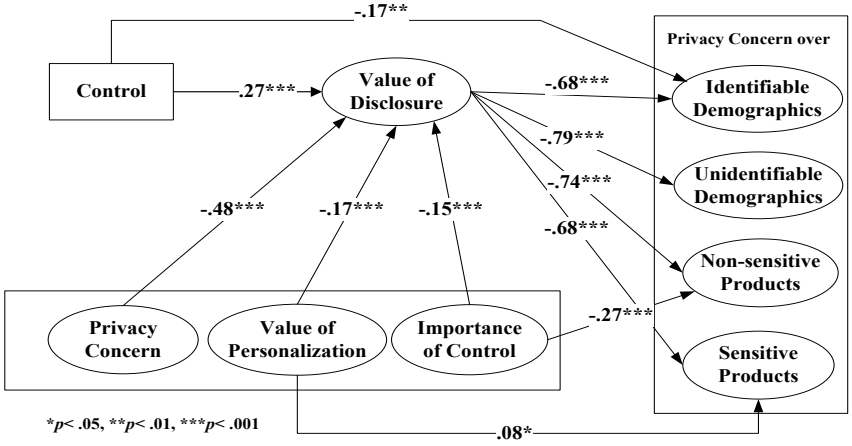


Figure 5. SEM explaining the psychological mechanism of privacy concerns

In sum, our findings suggest that different types of user information in a recommender system should be treated differently depending on the degree of privacy concern they may trigger. A user control mechanism is effective in reducing the concern regarding implicit data input only. In addition, control influences privacy concerns about both demographic and product-related information by way of users' perceived value of information disclosure.

5. DISCUSSION AND CONCLUSIONS

In this section, we provide interpretations of the study's main findings, present design suggestions for recommender systems, and then discuss the limitations and directions for future work.

5.1 Interpretation of Results and Design Implications

5.1.1 User Data Categorization and Sensitivity Ranking

The findings of our study suggest that users' online information is multi-dimensional regarding privacy concerns, especially in a recommender context. Although this seems self-explanatory, it is often neglected in privacy research and recommender system design. Specifically, demographic information that is frequently required for online service registration can be divided into two categories: unidentifiable information and identifiable information. Unidentifiable information consists of items describing one's personal attributes (e.g., age, gender) that cannot be used to uniquely pinpoint the individual, whereas identifiable information is more accurate in pointing to the individual's identity exclusively (e.g., phone numbers, email addresses). People are significantly more concerned about the recommender system accessing their identifiable information than their unidentifiable information. In a similar manner, product items can be broadly grouped into non-sensitive types and sensitive types. Users are significantly more worried about their previous purchases of sensitive products (e.g., adult diapers, HIV tests) being accessed for personalized recommendations than they are about their previous purchases of non-sensitive products (e.g., jewelry and shoes).

These item-based analyses and categorizations provide a relative information-ranking system in terms of privacy concern in recommender systems, thus refining existing research on general privacy concern about user information. Although a few previous studies have also identified specific information items that vary in sensitivity in recommender systems [27, 47], the current categorization extended prior research by extracting new factors, which can be used as a reference in future studies and designs. These new factors suggest that recommender system designers should treat users' information discriminatively and strategically based on their levels of sensitivity for pattern prediction and personalized recommendations. Algorithm developers should be well aware of what information users are more hesitant to disclose, so as to adjust the degree of information tracking and use, as well as to provide appropriate coping strategies. In line with the "privacy-personalization trade-off," unsolicited access to users' sensitive information may trigger severe privacy concerns that could affect users' overall experiences [28]; therefore, identifiable and sensitive data should be more cautiously handled in exchange for prediction accuracy. As a design suggestion, recommender systems should introduce user control or privacy assurance mechanisms to help alleviate users' privacy concerns. Also, user data with different sensitivity levels (e.g., identifiable vs. unidentifiable information) can be potentially protected with different levels of privacy remedies.

5.1.2 Effectiveness of User Control and Data Input Type

We also showed that the presence of a user control mechanism over information disclosure greatly impacts users' privacy concerns in a recommender system, which is consistent with previous findings [23, 24]. For demographic information, user control significantly lowered privacy concerns. However, for product-related information, such effects pertain to non-sensitive products only and are significantly moderated by data input type (i.e., explicit vs. implicit). When personalized recommendations are provided based on one's purchase history (i.e., implicit input), users tend to feel concerned about what they have bought when they have no control, but they feel significantly more relieved if they have control over information access by the service provider. This may be due to a sense of intrusiveness; implicit data input is often unsolicited, so users do not always expect that such information will be used for recommendation purposes. Affording users control over information release would not only allow users to modify their privacy settings and gain a sense of autonomy, but also help them predict what information might be at risk, thereby reducing the concern level resulting from uncertainty.

However, if users are explicitly told to rate the products they have purchased before (i.e., explicit input), the control mechanism does not help much in alleviating their concerns (Figure 3). Even though the control mechanism allows users to manage what information could be accessed by the recommender system, it seems that the control mechanism works for implicit data input rather than explicit data input. As discussed, implicit data input can trigger a sense of intrusiveness because records are often traced without users' permissions. On the other hand, explicit data input (i.e., product rating) is initiated by users, themselves, so users are already imbued with a sense of competency; because of this, an extra control mechanism would probably not change their perceptions. If users felt concerned about expressing their opinions and exposing their preferences, they would be unlikely to rate the products in the first place.

Furthermore, this intriguing interaction effect exists for non-sensitive products, but not for sensitive products. It could be that users are generally confident in protecting information related to non-sensitive products they have purchased, and the addition of a control mechanism further strengthens this confidence. However, when it involves sensitive products, users become much more cautious that their concern level may reach a "ceiling effect." Therefore, neither data input type nor the presence of control can alleviate the heightened concern levels.

This is the most intriguing finding of the current study, which casts doubts on ongoing efforts to embed user control in all recommender systems. The current study suggests that, for operations that do not require users' conscious attention and actions (e.g., tracking and analyzing their purchase history), an active control mechanism is needed to overcome perceived

intrusiveness and privacy concerns. However, users are already empowered with deliberation in an explicit rating situation, thus the extra control could seem redundant. Also, a control mechanism may only be convincing enough to protect information about non-sensitive products. As a design implication, a user control mechanism may not be as effective for recommendations relying on explicit data input, compared to those based on implicit user data input. Additionally, users seem to have persistent concerns about previously purchased sensitive products, and this cannot be easily mitigated by control mechanisms. There also is an asymmetric information problem between the service provider and the user—a lack of awareness could be another cause of the current finding. That is, users might not be aware of what companies can do with their non-sensitive information. Increasing the awareness level may boost the effectiveness of user control mechanisms. Therefore, system designers should carefully weigh the advantages and disadvantages of a control mechanism in addressing privacy concerns depending on the data input type, data sensitivity levels, and the existence of an awareness mechanism.

5.1.3 Psychological Mechanism of Privacy Concern

This study also proposed and tested a conceptual model for demonstrating the underlying psychological mechanisms of privacy concerns in a recommender system. Our findings showed that, after controlling for individual differences, users' perceived value of information disclosure explains how user control affects privacy concerns. In the current study, perceived value of information disclosure is measured based on the privacy calculus model, representing a trade-off between perceived benefits gained from personalized recommendations and risks of privacy invasion. Our results suggest that the mere mention of a control mechanism in the recommender system scenarios can elevate perceived value of information disclosure. This is likely because the addition of control boosts users' perceived value of the entire system, so users are more confident about trading in their privacy for personalized services. This heightened perceived value leads to lesser privacy concerns. Because perceived value often comes from perceived usefulness and effectiveness of the system [25], recommender system designers should focus on these aspects to improve users' psychological evaluation of the system so as to conquer privacy concerns. This is yet another motivation for designers to strive for a better recommender system with efficient functionality.

5.2 Limitations and Future Research

Although the current scenario-based design has its merits in many aspects, especially in an exploratory study, our manipulation and setting of the main constructs (e.g., control, data input) relied solely on users' assumptions and imaginations as instructed by our study descriptions. Participants may have had a different impression and evaluation of a recommender system if they could interact with a real interface. Their concerns over various information types also depended on a hypothetical picture of what they had previously purchased from Amazon.com. Therefore, the scenario-based design may lack external validity. Future research could implement a real interface prototype based on our preliminary findings, examine users' real behaviors (e.g., purchasing, rating) in a natural setting over a longitudinal period, and then measure their privacy concern levels. In addition, apart from perceived value of disclosure, other psychological mechanisms of privacy concerns in recommender systems should also be explored.

6. REFERENCES

[1]Acquisti, A. and Gross, R. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In Proceedings of the Privacy enhancing technologies. 36-58.

[2]Arlein, R.M., Jai, B., Jakobsson, M., Monrose, F., and Reiter, M.K. 2000. Privacy-preserving global customization. In Proceedings of the 2nd ACM conference on Electronic commerce. 176-184.

[3]Awad, N.F. and Krishnan, M. 2006. The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to be Profiled Online for Personalization. MIS quarterly, **30**(1).

[4]Bansal, G., Zahedi, F., and Gefen, D. 2010. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. Decision Support Systems, **49**(2): 138-150.

[5]Belanger, F., Hiller, J.S., and Smith, W.J. 2002. Trustworthiness in electronic commerce: the role of privacy, security, and site attributes. The Journal of Strategic Information Systems, **11**(3): 245-270.

[6]Bennett, J. and Lanning, S. 2007. The netflix prize. In Proceedings of the KDD cup and workshop. 35.

[7]Berendt, B. and Teltzrow, M. 2005. Addressing users' privacy concerns for improving personalization quality: Towards an integration of user studies and algorithm evaluation, in Intelligent Techniques for Web Personalization. Springer. 69-88.

[8]Burke, R. 2002. Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, **12**(4): 331-370.

[9]Chellappa, R.K. and Sin, R.G. 2005. Personalization versus privacy: An empirical examination of the online consumer's dilemma. Information Technology and Management, **6**(2-3): 181-202.

[10]Chen, L. and Pu, P. 2012. Critiquing-based recommenders: survey and emerging trends. User Modeling and User-Adapted Interaction, **22**(1-2): 125-150.

[11]Cranor, L.F. 2004. I didn't buy it for myself, in Designing personalized user experiences in eCommerce. Springer. 57-73.

[12]Culnan, M.J. 1993. " How Did They Get My Name?": An Exploratory Investigation of Consumer Attitudes Toward Secondary Information Use. MIS quarterly, **17**(3).

[13]Culnan, M.J. 2000. Protecting privacy online: Is self-regulation working? Journal of Public Policy & Marketing, **19**(1): 20-26.

[14]Culnan, M.J. and Armstrong, P.K. 1999. Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. Organization science, **10**(1): 104-115.

[15]Culnan, M.J. and Bies, R.J. 2003. Consumer privacy: Balancing economic and justice considerations. Journal of social issues, **59**(2): 323-342.

[16]Edwards, S.M., Li, H., and Lee, J.-H. 2002. Forced exposure and psychological reactance: Antecedents and

consequences of the perceived intrusiveness of pop-up ads. Journal of Advertising, **31**(3): 83-95.
[17]Finch, J.F. and West, S.G. 1997. The investigation of personality structure: Statistical models. Journal of Research in Personality, **31**(4): 439-485.
[18]Friedman, B., Khan Jr, P.H., and Howe, D.C. 2000. Trust online. Communications of the ACM, **43**(12): 34-40.
[19]Gena, C., Brogi, R., Cena, F., and Vernero, F. 2011. The impact of rating scales on user's rating behavior, in User Modeling, Adaption and Personalization. Springer. 123-134.
[20]Hauser, J.R., Urban, G.L., Liberali, G., and Braun, M. 2009. Website morphing. Marketing Science, **28**(2): 202-223.
[21]Hayes, A.F. 2008. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach: Guilford Press.
[22]Heitmann, B., Kim, J.G., Passant, A., Hayes, C., and Kim, H.-G. 2010. An architecture for privacy-enabled user profile portability on the Web of Data. In Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems. 16-23.
[23]Kay, J. and Kummerfeld, B. 2006. Scrutability, user control and privacy for distributed personalization. In Proceedings of the CHI 2006 Workshop on Privacy-Enhanced Personalization. 21-22.
[24]Kay, J., Kummerfeld, B., and Lauder, P. 2002. Personis: a server for user models. In Proceedings of the Adaptive Hypermedia and Adaptive Web-Based Systems. 203-212.
[25]Kim, H.-W., Chan, H.C., and Gupta, S. 2007. Value-based adoption of mobile internet: an empirical investigation. Decision Support Systems, **43**(1): 111-126.
[26]Kittur, A., Chi, E.H., and Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. In Proceedings of the Conference on Human Factors in Computing Systems 2008. 453-456.
[27]Knijnenburg, B.P. and Kobsa, A. 2013. Making decisions about privacy: information disclosure in context-aware recommender systems. ACM Transactions on Interactive Intelligent Systems (TiiS), **3**(3): 20.
[28]Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., and Newell, C. 2012. Explaining the user experience of recommender systems. User Modeling and User-Adapted Interaction, **22**(4-5): 441-504.
[29]Komiak, S.Y. and Benbasat, I. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. MIS quarterly: 941-960.
[30]Koren, Y., Bell, R., and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. Computer, **42**(8): 30-37.
[31]Lampe, C., Ellison, N.B., and Steinfield, C. 2008. Changes in use and perception of Facebook. In Proceedings of the 2008 ACM conference on Computer supported cooperative work. 721-730.
[32]Li, T. and Unger, T. 2012. Willing to pay for quality personalization? Trade-off between quality and privacy. European Journal of Information Systems, **21**(6): 621-642.
[33]Lin, J., Amini, S., Hong, J.I., Sadeh, N., Lindqvist, J., and Zhang, J. 2012. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing. 501-510.
[34]Malhotra, N.K., Kim, S.S., and Agarwal, J. 2004. Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. Information Systems Research, **15**(4): 336-355.
[35]Margulis, S.T. 2003. Privacy as a social issue and behavioral concept. Journal of social issues, **59**(2): 243-261.
[36]Metzger, M.J. 2006. Effects of site, vendor, and consumer characteristics on web site trust and disclosure. Communication Research, **33**(3): 155-179.
[37]Milne, G.R. and Boza, M.-E. 1999. Trust and concern in consumers' perceptions of marketing information management practices. Journal of Interactive marketing, **13**(1): 5-24.
[38]Mount, M.K., Barrick, M.R., Scullen, S.M., and Rounds, J. 2005. Higher‐order dimensions of the big five personality traits and the big six vocational interest types. Personnel Psychology, **58**(2): 447-478.
[39]Peppers, D., Rogers, M., and Dorf, B. 1999. Is your company ready for one-to-one marketing. Harvard Business Review, **77**(1): 151-160.
[40]Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.-P., and Jonker, C.M. 2012. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. User Modeling and User-Adapted Interaction, **22**(4-5): 357-397.
[41]Sheehan, K.B. and Hoy, M.G. 2000. Dimensions of privacy concern among online consumers. Journal of Public Policy & Marketing, **19**(1): 62-73.
[42]Shneiderman, B. 2000. Designing trust into online experiences. Communications of the ACM, **43**(12): 57-59.
[43]Smith, H.J., Milberg, S.J., and Burke, S.J. 1996. Information Privacy: Measuring Individuals' Concerns About Organizational Practices. MIS quarterly, **20**(2).
[44]Sundar, S. and Marathe, S. 2006. Is it tailoring or is it agency? Unpacking the psychological appeal of customized news. In Proceedings of the the 89th Annual Convention of the Association for Education in Journalism and Mass Communication.
[45]Sundar, S.S., Kang, H., Wu, M., Go, E., and Zhang, B. 2013. Unlocking the privacy paradox: do cognitive heuristics hold the key? In Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems. 811-816.
[46]Toch, E., Wang, Y., and Cranor, L.F. 2012. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. User Modeling and User-Adapted Interaction, **22**(1-2): 203-220.
[47]Tsai, J.Y., Egelman, S., Cranor, L., and Acquisti, A. 2011. The effect of online privacy information on purchasing behavior: An experimental study. Information Systems Research, **22**(2): 254-268.
[48]Unni, R. and Harmon, R. 2007. Perceived effectiveness of push vs. pull mobile location based advertising. Journal of Interactive advertising, **7**(2): 28-40.

[49]Westin, A.F. 1968. Privacy and freedom. Washington and Lee Law Review, **25**(1): 166.

[50]Xiao, B. and Benbasat, I. 2007. E-commerce product recommendation agents: use, characteristics, and impact. MIS quarterly, **31**(1): 137-209.

[51]Xu, H., Dinev, T., Smith, J., and Hart, P. 2011. Information Privacy Concerns: Linking Individual Perceptions with Institutional Privacy Assurances. Journal of the Association for Information Systems, **12**(12).

[52]Xu, H., Luo, X.R., Carroll, J.M., and Rosson, M.B. 2011. The personalization privacy paradox: an exploratory study of decision making process for location-aware marketing. Decision Support Systems, **51**(1): 42-52.

[53]Xu, H., Teo, H.-H., Tan, B.C., and Agarwal, R. 2012. Research Note-Effects of Individual Self-Protection, Industry Self-Regulation, and Government Regulation on Privacy Concerns: A Study of Location-Based Services. Information Systems Research, **23**(4): 1342-1363.

[54]Zhang, B., Wu, M., Kang, H., Go, E., and Sundar, S.S. 2014. Effects of security warnings and instant gratification cues on attitudes toward mobile websites. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems. 111-114.

7. APPENDIX

A. Scenarios

Condition			Scenario
Without Control	Demographics		How **CONCERNED** would you feel if Amazon.com accessed the following information about you in return for personalized recommendations, without asking you first?
	Products	Implicit Input	Suppose **YOU HAVE PURCHASED** the following items from Amazon.com. Please indicate how **CONCERNED** you would feel for Amazon.com to access your purchase history of each of the following items in return for personalized recommendations.
		Explicit Input	Suppose you **HAVE PURCHASED** the following items from Amazon.com. Please indicate how **CONCERNED** you would feel to provide your **RATINGS** of the items to Amazon.com in return for personalized recommendations.
With Control	Demographics		Suppose you **HAVE CONTROL** over the extent to which Amazon.com can access your personal information. With such control, how **CONCERNED** would you feel having the following information about you stored on Amazon.com?
	Products	Implicit Input	Suppose you **HAVE CONTROL** over the extent to which the following items **IN YOUR PURCHASE HISTORY** can be accessed by Amazon.com. With such control, please indicate how **CONCERNED** you would feel having each of the items in your purchase history on Amazon.com.
		Explicit Input	Suppose you **HAVE CONTROL** over the extent to which **YOUR RATINGS** of the following items can be accessed by Amazon.com. With such control, please indicate how **CONCERNED** you would feel to RATE each of the items in return for personalized recommendations.

B. Information Items

Related questions (see Appendix A) were answered on a scale of "1 = Not Concerned at All" to "7 = Extremely Concerned."

Demographic Information	Product-Related Information
1. Gender	1. Textbooks
2. Age	2. Digital Games
3. Education	3. Jewelry
4. Race	4. Furniture
5. Relationship status	5. Snack Food
6. Technology use	6. Flowers
7. Email address	7. Shoes
8. Phone number	8. Laptop
9. Credit card number	9. Lingerie
10. Social security number	10. Condoms
11. Date of birth	11. Lubricant
12. Name	12. Book – Depression
13. Home address	13. Weight Loss Products
14. Company	14. Pregnancy Test
15. Interest areas	15. Book – Bankruptcy
16. Field of work	16. Fertilizer
17. Household income	17. Adult Diapers
18. Location	18. Hunting Knife
19. Calendar data	19. Cigarettes
20. Web browsing history	20. Bottle of Peroxide
21. IP address	21. Sex Toys
	22. HIV Test
	23. Pornographic DVD
	24. STD Medication
	25. Bulletproof Jacket
	26. Book - Bomb-Making

C. Measurements

These measures were all based on a scale of "1 = Strongly Disagree" to "7 = Strongly Agree" unless otherwise noted.

Perceived Value of Information Disclosure

1. I think my benefits gained from using Amazon.com's service can offset the risks of my information disclosure.
2. The value I gain from using Amazon.com's service is worth the information I give away.
3. I think the risks of my information disclosure will be greater than the benefits gained from using Amazon.com's service.

Trust (Please indicate how well each of the following adjectives describes Amazon.com.)

1. Reliable
2. Trustworthy
3. Dependable
4. Honest
5. Fair
6. Exploitative (reverse coded)

Perceived Value of Online Personalization

1. I value web pages that are personalized for the device (e.g., computer, tablet, mobile phone, etc.), browser (e.g., Internet Explorer, Firefox, Chrome, etc.) and operating system (e.g. Windows, Mac OS, Unix) that I use.
2. I value websites that are personalized for my usage experience preferences.
3. I value websites that acquire my personal preferences and personalize the services and products themselves.
4. I value goods and services that are personalized based on information that is collected automatically (e.g., IP address, web browsing history) but cannot identify me as an individual.
5. I value goods and services that are personalized based on information that I have voluntarily given out (e.g., age, household income, field of work) but cannot identify me as an individual.
6. I value goods and services that are personalized on information I have voluntarily given out and can identify me as an individual (e.g., name, address, credit card number).

Importance of Control

1. It is important for me to restrict Amazon.com's use of a specific type of information for personalized recommendations.
2. It is important for me to control Amazon.com's access of a specific type of information for personalized recommendations.
3. It is important for me to control the amount of information accessed by Amazon.com for personalized recommendations.

General Privacy Concern

1. I am sensitive about giving out information regarding my preferences.
2. I am concerned about anonymous information (information collected automatically but cannot be used to identify me, such as my computer, network information, operating system, etc.) that is collected about me.
3. I am concerned about how my personally un-identifiable information (information that I have voluntarily given out but cannot be used to identify me, e.g., age, gender, field of work, etc.) will be used by firms.
4. I am concerned about how my personally identifiable information (information that I have voluntarily given out AND can be used to identify me as an individual, e.g., name, home address, credit card number, etc.) will be used by firms.

Demographics (These were all posed as multiple-choice questions)

1. What is your age?
2. What is your gender?
3. What was the highest level of education you have received?
4. What racial group do you belong to?

Behavioral Experiments Exploring Victims' Response to Cyber-based Financial Fraud and Identity Theft Scenario Simulations

Heather Rosoff
Sol Price School of Public Policy
University of Southern California
rosoff@usc.edu

Jinshu Cui
Department of Psychology
University of Southern California
jinshucu@usc.edu

Richard John
Department of Psychology
University of Southern California
richardj@usc.edu

ABSTRACT

We conducted two scenario-simulation behavioral experiments to explore individual users' response to common cyber-based financial fraud and identity theft attacks depend on systematically manipulated variables related to characteristics of the attack and the attacker. Experiment I employed a 4 by 2 between-groups factorial design, manipulating attacker characteristics (individual with picture vs. individual vs. group vs. unknown) and attack mode (acquiring a bank database vs. obtaining personal bank account information) in response to a bank letter scenario notifying respondents of a data breach. Respondents' positive and negative affect, perceived risk, behavioral intention and attitude towards the government's role in cyber security were measured. Results suggest that respondents experienced greater negative affect when the attacker was an individual, as well as experienced more positive affect when the attack target was an individual bank account. In addition, a picture of an individual attacker increased intended behavioral changes and expectations of the bank to manage the response in the bank database attacks only. Experiment II utilized a 4 by 3 between-groups factorial design, manipulating attacker motivation (fame vs. money vs. terrorism vs. unknown) and attack resolution status (resolved vs. still at risk vs. unknown) in response to an identity theft scenario that evolves over four time points. In this experiment, respondents' affect, perceived risk and intended short- and long-term behavior were measured at each time point. Results suggest that respondents reported less perceived risk when the attacker's motivation was to fund terrorism. Respondents also reported lower negative affect and lower perceived risk when the identity theft case was reported as resolved. Respondents also were more willing to pursue long-term behavior changes when the attack outcome was still at risk or unknown. In both experiments, respondents' sex and age were related to affect, risk perception, and behavioral intentions. The paper also includes discussion of how further understanding of individual user decision making informs policy makers' design and implementation of cyber security policies related to credit fraud and identity theft.

1. INTRODUCTION

With the advent of the information age, cyber attacks have exploded as a major concern. As stated by the Officer-in-charge at the United National Interregional Crime and Justice Research Institute (UNICRI), "The likelihood of suffering from a real crime, like being robbed in the street, is now smaller than the possibility of suffering a virtual crime, such as an online identity theft or a credit card fraud." Individual users' decision making is critical to determining whether a cyber attack can be committed and what the extent of that damage might be (Rosoff, Cui, & John, 2013). This is complicated by the information asymmetry between the attackers and individual users. With limited information as to the causes and consequences of cyber threats, individual users often trigger attacks unintentionally and consequently react poorly and suffer from severe outcomes. While the characteristics and motivations of attackers have been investigated thoroughly by defenders to better understand how to detect threats and protect cyber systems (D'Amico, Whitley, Tesone, O'Brien, & Roth, 2005; Liu, Yu, & Mylopoulos, 2003; Nykodym, Taylor, & Vilela, 2005), there is limited research on how information about attackers influences individual users' emotional, cognitive and behavioral responses to cyber threats.

This paper reports the results of two scenario-based experiments of a cyber-based financial fraud or identity theft attack. These experiments utilize a scenario simulation methodology that includes an experimental manipulation, instead of the traditional survey-based scenario, as well as stimulus material to enhance the scenario's realism. More specifically, in Experiment I we explored whether attacker characteristics and attack mode influenced the victim's reaction and behavioral response to a data breach at their bank. In Experiment II, we assessed whether the attacker's motivation and the resolution status of the attack affected the victim's emotional, cognitive, and behavioral response for an identity theft case. We believe the use of narrative scenarios and images are more compelling and concrete to respondents, and increase the likelihood of obtaining valid responses compared to less concrete scenario stimuli. Furthermore, in both scenario simulations, all but the manipulated variables are held constant so that any significant findings can be attributed to the manipulated variables.

In Experiment I, we hypothesized that attacker characteristics, specifically those accompanied by a photograph, would decrease feelings of vulnerability and result in fewer behavioral changes in response to the cyber-based data breach at the bank (financial fraud). This hypothesis follows from construal theory (Trope and Liberman, 2010); pictures are more concrete representations, resulting in a lower level of construal, compared to words which are more abstract and distant representations associated with higher level construal. This finding has been reported in the disaster literature and has shown that images have the potential to lower negative affect and perceived risk (Peters and Slovic, 1996; Leiserowitz, 2006). Furthermore, the direction of behavioral decision making, with respect to level of involvement in response efforts, tends to coincide with affective and risk reactions; lower perceived risk and negative affect more often predict more moderated behavioral changes in response to an event (Terpstra, 2011).

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9-11, 2014, Menlo Park, CA.

Also for Experiment I, we anticipated that there would be some influence of attack mode on affective reactions and behavioral responses to the bank data breach. Crime research has shown that personal victims of crime experience increased fear and vulnerability that translates into a greater willingness to adopt crime reduction measures. This is compared to widespread neighborhood crime where collectively victims also experience increased feelings of vulnerability, yet their willingness to act is moderated by their expectation of local officials to be proactively involved in the response (Skogan and Maxfield, 1982; Norris et al., 2008; L.W., 2012). We anticipated that in the cyber context, individual victims of an attack on a personal bank account or group victims of an attack on a bank database also would have a negative reaction to the event. We expected that for the victims of the personal bank account attack this would lead to more proactive efforts to resolve the consequences associated with the data breach compared to the database victims.

In Experiment II we anticipated that the attacker's motivation would depend on the perceived psychological distance from the cyber-based identity theft case, and in turn, this would influence users' perceived risk and decision making. This expectation is also based on construal theory which suggests that the more distant an object is from the individual, the more abstract it will be thought of, while the closer the object is, the more concretely it will be thought of (Trope & Liberman, 2003, 2010; Trope, Liberman, & Wakslak, 2007). In the cyber context, we expected that the more concrete the attacker motivation, the greater the perceived risk of identity theft.

Also in Experiment II we explored the extent to which the resolution of the identity theft case influenced victim's thinking and behavioral reactions to the attack. We hypothesized that the level of uncertainty associated with an unresolved or unknown outcome would threaten victims' sense of control, resulting in increased negative affect and heightened risk perceptions (Slovic, Fischhoff and Lichtenstein, 1980; Vlec and Stallen, 1980) Furthermore, respondents are believed to perceive the unresolved and unknown identity theft case outcomes as putting them in harm's way, which also is a determinant of elevated behavioral responses (Slovic, Fischhoff and Lichtenstein, 1984; Slovic, 1987).

Lastly, we considered how demographic variables affect the strength and/or the direction of the relationship between the manipulated variables, attacker characteristics, attacker motivation, attack mode, and attack resolution status, and the dependent variables, affect, perceived risk and behavioral intention. For example, one possibility is that the perceived risks posed by financial fraud or identity theft tend to be judged lower by men than women (Garbarino et al., 2004; Bhatnager and Misra, 2000); consequently, women are expected to have a stronger desire than men to modify their cyber behavior. Another possibility is that the reliance on a third party to assist in the necessary behavior change in response to financial fraud or identity theft would be less for younger users because they are more familiar and comfortable with the nuances of internet security options. Overall, we anticipated that there would be some difference in the patterns of response as a function of sex and age for Experiment I, and sex for Experiment II.

The next section of this article describes the methods, results, and a brief discussion for Experiment I, and Section 3 describes the methods, results, and a brief discussion for Experiment II. The paper closes with a discussion of study limitations and how these results have the potential to enhance and improve cyber security by taking into account end-user decision making.

2. EXPERIMENT I

2.1 Methods

In August of 2013, we conducted an experiment involving a cyber-based bank attack with two manipulated variables, attacker characteristics and attack mode, to evaluate individual's emotional response, perceived risk, and behavioral intention in response to the event. The bank data breach scenario was developed to capture a common financial fraud event that significantly affects individual users. More specifically, the dependent variables focused on individuals' positive and negative feelings about the event, the perceived risk to financial security and the likelihood of a second event, and decision making related to banking, ranging from relying on the bank to manage the attack response versus discontinuing all banking activity.

Table 1. Scenario and Manipulations (Experiment I)

Manipulations	Scenario			
	August 2, 2013 Dear Valued Customer, We are writing to notify you that two days ago,			
Attack mode	there was an unauthorized attempt to withdraw all of your current funds. *(personal)*	there was an unauthorized breach into our customer information center, which stores credit card and personal information for all 10 million of our clients *(database)*.		
Attacker characteristics	As of now, we know an individual online hacker is responsible for the breach into your account. The hacker acted alone in carrying out the attack. *(individual)*	As of now, we know a hacking group is responsible for the breach into your account. An organization of hackers coordinated the attack. *(group)*	As of now, we do not know if a hacking group or an individual hacker is responsible for the breach. *(unknown)*	As of now, we know the individual online hacker pictured below is responsible for the breach into your account. The hacker acted alone in carrying out the attack. *(individual with picture)*
	We are working with law enforcement officials and regret any concern or inconvenience this incident may have caused you. We will keep you informed as we make progress in his capture. Kindest Regards, Your Bank			

2.1.1 Design Overview

A 4 (attacker characteristics) by 2 (attack mode) between-groups factorial design was used to explore responses to a bank letter notifying respondents of a data breach. Each respondent was randomly assigned to one of eight conditions. The four attacker characteristics are (1) individual with picture, (2) individual, (3) group, and (4) unknown; the two attack modes are (1) acquiring a bank database and (2) obtaining personal bank account information. The experiment was submitted to the University of Southern California's Institutional Review Board (IRB) and the IRB determined that study qualified for Exempt, Category 2 research.

The experiment opened with respondents first providing demographic information (sex and age) and answering a series of questions regarding their previous online experience. They were then presented with the bank notification letter. The content of the notification and manipulations contained is provided in Table 1.

After reading the bank notification, respondents were asked to evaluate their negative affect, positive affect, cyber risk perception, threat belief, intended behavioral response, and attitudes toward the role of government in preventing cyber attacks.

2.1.2 Measures

Respondents' current feelings, risk perception, behavioral intention and attitude towards the government's role in cyber security were measured following receipt of a bank notification alerting the respondent to the cyber attack. Details of the items in each measure are included in Table 2.

Affect. The Positive Affect Negative Affect Scale (PANAS) was included to measure self-reported emotion (Watson, Clark, & Tellegen, 1988). The version used was an abbreviated 10-item PANAS (Rosoff, Siko, John, & Burns, 2013). Each affect item was rated from 1(not at all) to 5 (extremely). Principal axis factoring was performed on the 10-item PANAS and two factors were extracted. Items were internally consistent with Cronbach's alphas = .94 and .84 for negative and positive affect, respectively.

Risk Perception. Respondents also were asked to estimate personal financial safety using a scale from 0 (not at all risky) to 10 (extremely risky), vulnerability to identity theft using a scale from 0 (not at all vulnerable) to 10 (extremely vulnerable), likelihood of an attempted second attack using a scale from 0% (not at all likely) to 100% (very likely), and likelihood of a successful second attack using a scale from 0% (not at all likely) to 100% (very likely). The scores for the first two items were multiplied by 10 to equal the ranges of the likelihood items. Principal axis factoring was performed and one factor was extracted. Items were internally consistent with a Cronbach's alpha = .83.

Behavioral Intention. Respondents assessed their intended behavior on a 5-point scale (1=strongly disagree to 5=strongly agree). From the six behavioral intention questions, two factors were extracted. The first factor, moderate behavioral intention, captured expectations relative to the bank's response to the event. This included "get credit checked", "expect bank to enhance security", and "expect bank to reimburse" with a Cronbach's alpha = .63 and rotated loadings all above .68. The second factor, severe behavioral intention, addressed behavioral decisions related to discontinuing the use of financial services. This factor included "no longer online bank", "cancel credit cards", and "discontinue all online financial activities" with a Cronbach's alpha = .75 and rotated loadings all above .69.

Attitude towards the Government's Role in Cyber Security. Respondents evaluated their attitude towards the government's role in online protection on a 5-point scale (again 1=strongly disagree to 5=strongly agree) for 4 items listed in Table 2. The four items were internally consistent with a Cronbach's alpha = .71.

Table 2. Measures of Experiment I

Scales	Items
Negative affect	distressed, afraid, upset, nervous, scared
Positive affect	enthusiastic, inspired, strong, determined, active
Risk perception	(1) What do you believe the risk is to your personal financial safety?
	(2) How vulnerable do you believe you are to becoming a victim of identity theft?
	(3) What do you believe to be the likelihood of an attempted second cyber attack on your bank?
	(4) What do you believe to be the likelihood of a successful second cyber attack on your bank?
Intended behavior	(1) I would start using another bank.
	(2) I would no longer online bank.
	(3) I would get my credit checked.
	(4) I would cancel my credit cards.
	(5) I would expect my bank to enhance its security.
	(6) I would expect my bank to reimburse me for any fraudulent charges on my account.
	(7) The hacker(s) responsible for the cyber attack described should go to jail.
	(8) I would discontinue all online financial activities.
Attitudes toward government role in cyber security	(1) I am not willing to give up some of my privacy for greater online protection.
	(2) The government needs to increase its Internet security initiatives.
	(3) I don't mind if the government has access to my personal information in order to increase security.
	(4) I am not worried about cyber attacks on the American government.

Table 3. Demographic Information and Cyber-related Experience

Variables (N = 239)	Response Category	Number and Percentage
Do you shop online?	Yes	235 (98.3%)
	No	4 (1.7%)
	I don't know	0 (.0%)
Do you bank online?	Yes	222 (92.9%)
	No	16 (6.7%)
	I don't know	1 (.4%)
Has your identity ever been stolen?	Yes	15 (6.3%)
	No	214 (89.5%)
	I don't know	10 (4.2%)
Has your credit card ever been stolen?	Yes	51 (21.3%)
	No	186 (77.8%)
	I don't know	2 (.8%)
Have you been trained in Internet security either independently or by your employer?	Yes	54 (22.6%)
	No	182 (76.2%)
	I don't know	3 (1.3%)
Sex	Male	136 (56.9%)
	Female	103 (43.1%)
Age	18-25	68 (28.5%)
	26-30	50 (20.9%)
	31-35	37 (15.5%)
	36-40	25 (10.5%)
	41-45	21 (8.8%)
	46-50	10 (4.2%)
	51-55	9 (3.8%)
	56-60	10 (4.2%)
	61-65	5 (2.1%)
	66+	4 (1.7%)

2.1.3 Respondents

The experiment was hosted on Qualtrics.com and respondents were recruited from Amazon Mechanical Turk (AMT). Two-hundred and forty-three adult respondents participated and were paid $0.55 for their participation. Four of the 243 respondents were removed for answering the attention check question incorrectly. Two-hundred and thirty-nine respondents were included in the analysis. The number of respondents assigned to each of the eight design conditions ranged from 29 to 31. The median time for completion was 6 minutes. Table 3 provides demographic information and a summary of cyber-related experience for the respondents.

Composite scores were calculated across items using equal weighting for the six dependent variables: negative affect, positive affect, risk perception, severe behavior, moderate behavior, and attitudes toward government role.

2.2 Results

Least squares regression was used to predict respondents' scores on the six dependent variables (positive affect, negative affect, risk perception, moderate behavioral intention, extreme behavioral intention and attitude towards the government's role in cyber security) from the two manipulated variables (attacker characteristics and attack mode), and respondent characteristics (sex and age). To fully examine the influence of attacker characteristics, three orthogonal contrasts were created and entered into the regressions as independent variables: (1) individual and individual with picture vs. group and unknown, (2) individual vs. individual with picture, and (3) group vs. unknown.

Results indicate that respondents' negative affect was significantly greater when the cyber attack was conducted by an individual attacker compared to an individual attacker with a picture (standardized $\beta = .135$, $t = 2.088$, $p = .038$, $R^2 = .068$). No significant difference was found between an individual attacker and an individual attacker with picture for reported positive affect, risk perception, intended behavior and attitudes toward the government. In addition, positive affect was found to be significantly influenced by the attacker's selected attack mode (standardized $\beta = .143$, $t = 2.275$, $p = .024$, $R^2 = .108$). Respondents experienced more positive affect when their personal account was directly attacked compared to a compromised bank database. Negative affect, risk perception, intended behavior and attitude towards the government were not significantly influenced by the attacker's selected attack mode.

A significant interaction between attacker characteristics and attack mode relative to respondents' expectations of bank services was also found (standardized $\beta = .137$, $t = 2.106$, $p = .036$, $R^2 = .044$). Interestingly, there was an expectation from all respondents that the bank would enhance its security in response to the security breach. Moreover, respondents had even higher expectations of the bank to resolve the cyber attack when the attack was targeted against their personal account versus the bank's database, independent of the attacker's characteristics. In addition, when the attacker directly targeted only the personal account of the victim, the expectation for bank involvement was significantly greater when the individual attacker was presented with a picture compared to no picture. No significant interaction effect was found between the two manipulated variables for negative affect, positive affect, risk perception, intended behavior, and attitude towards the government. Moreover, emotional, cognitive, and behavioral reactions were found to not differ significantly between individual attacker and individual attacker with a picture vs. group and unknown attacker and between group vs. unknown attacker. Figure 1 displays the mean negative affect, positive affect and expectation of the bank/intent to continue bank service for different attacker characteristics and attack modes.

Lastly, regression results indicated that sex was predictive of negative affect (standardized $\beta = .165$, $t = 2.466$, $p = .014$), risk perception (standardized $\beta = .155$, $t = 2.301$, $p = .022$), and attitudes toward the role of government in preventing a cyber attack (standardized $\beta = .151$, $t = 2.324$, $p = .021$). Female respondents tended to experience more negative affect, perceive more risk, and were more likely to support the government's intervention. Sex was not significantly predictive of positive affect and behavioral intention. Age also was found to significantly predict positive affect (standardized $\beta = .278$, $t = 4.295$, $p < .001$) and attitudes toward the government's role in online protection (standardized $\beta = .189$, $t = 2.924$, $p = .004$). Older respondents tended to experience more positive affect and greater support for the government's intervention in online security. Age was not found to significantly predict negative affect, risk perception, and behavioral intention.

2.3 Discussion

The results of Experiment I suggest that respondents negative affect, positive affect and expectation of the bank's response (moderate behavioral intention) to the cyber-based bank data

breach were significantly influenced by the manipulation of attacker characteristics and attack mode. Consistent with our hypothesis, respondents appear to have experienced less negative affect because the picture is interpreted as a more concrete, less distant representation of the attacker. While traditional construal level theory research has found that more concrete objects are associated with greater negative affect, in the cyber attack context this pattern of results is reversed. This is because the perception of the attacker in cyber space is abstract and distant, resulting in a baseline of high negative affect. As such, as the attacker becomes more familiar and close through a picture, negative affect is shown to decrease.

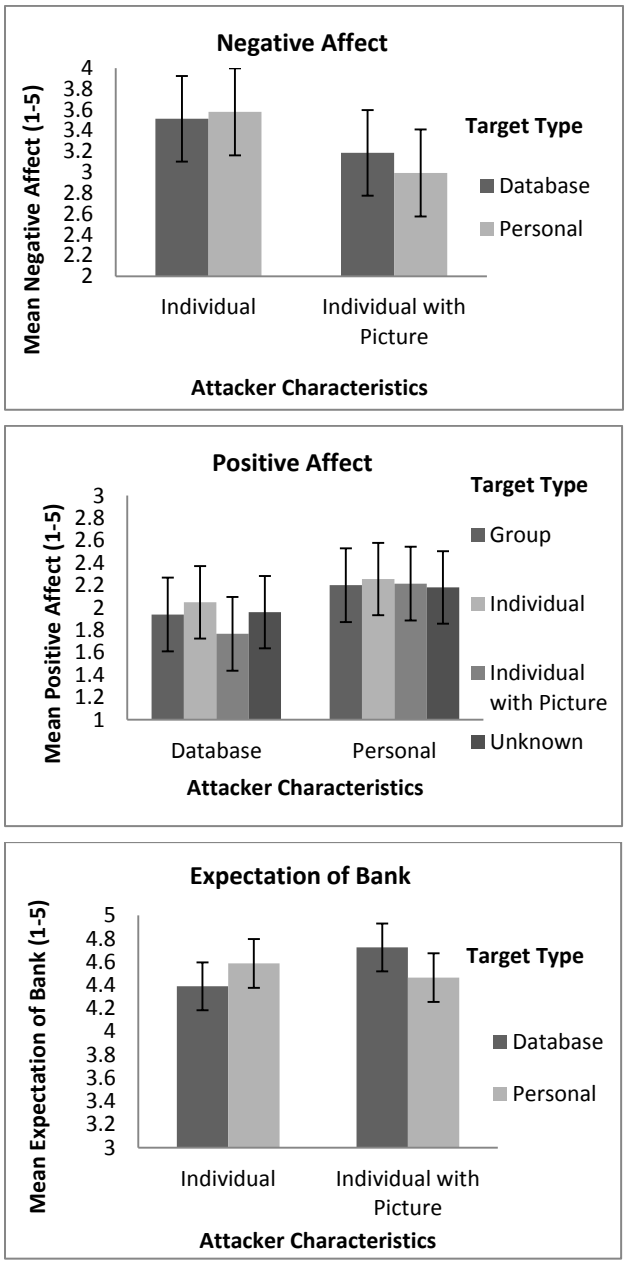


Figure 1. Mean negative affect, positive affect, moderate behavioral intention - to continue using banking services - for attacker characteristics and attack targets. Note: Error bars are +/- 2 SE.

We also found that respondents felt more enthusiastic, inspired, strong, determined, and active when only their personal account was victimized than when their bank's account database was compromised. As described above, positive affect includes words suggestive of the amount of energy one would expend in response to the cyber threat. As such, respondents expressed a greater desire to take action only when their personal account was hacked compared to victims of the database hack for which the responsibility to act tended to be diffused. One explanation is that the attack mode (in this case, the bank) is likely to take the lead in the response effort to protect again the potential cost of the attack to their reputation and profit/success. One might expect that banks make the needs of database victims a priority, which diffuses the desire to act between the attack mode owner and database members.

In addition, there was a significant interaction between attacker characteristics and attack mode. When a personal account was hacked, respondents were more likely to count on the bank if the picture of the attacker was presented. Conversely, when the bank database was compromised, respondents were indifferent to whether the picture of the attacker was provided.

Also, as anticipated, we found that female respondents experienced more negative affect, perceive more risk, and were more likely to support the government's intervention in online security. Disaster risk perception studies also have shown that risks tend to be judged higher by females (Kung and Chen 2012; Bourque et al. 2012) and that females tend to have a stronger desire to take preventative and preparedness measures compared with males (Ho et al. 2008; Cameron and Shah 2012). We also found an age effect suggesting that older respondents tended to experience more positive affect and in turn, were more likely to support government intervention in online security. Results related to the role of victim age in the crime and disaster literature have been conflicting (Hale, 1996; Fischhoff, 2003; Sjöberg, 2005; Henson, Reyns, & Fisher, 2013) and our findings reflect the perspective that there are significant age differences.

The overall policy implications of these findings depend on the financial institutions' respective objectives. If the ultimate goal is to calm bank members down following a cyber breach, as opposed to enhance their concern and increase their avoidance behavior, our findings suggest that sharing a photo has the potential to be helpful. However, if the financial institution is interested in having both the bank and bank members engage in protective behaviors, sharing a photo of the attacker does not appear to be the best tactic for encouraging member emotional investment in threat resolution and engagement in avoidance behavior. Gender and age findings further suggest that females and older respondents have the potential to be more inclined to support policy recommendations. However, additional research relative to specific policy compliance with a larger sample is needed for an assessment of the moderating effects of demographic variables. Overall, such variations in policy considerations allow for financial institutions to more effectively assess the trade-offs between the social impacts and costs associated with policy implementation.

3. EXPERIMENT II

3.1 Methods

In Experiment II, we continued to manipulate attacker attributes. Particularly, we manipulated attacker's motivations to commit a

cyber attack and measured the impact on respondents' emotional reactions, perceived risk and decision making. We also continued to explore the impact of financial fraud in the context of an identity theft scenario that evolves over four time points (compared to the bank data breach scenario described at a single time point). Similar to the bank data breach scenario, identity theft has the potential to present serious inconveniences for the victim. The time and effort a victim might have to spend responding to and resolving an identity theft case could be substantial. Therefore, we also manipulated the level of resolution associated with the outcome of the identity theft scenario. In order to ensure that the significant findings were attributed to the manipulated variables, attacker motivations and resolution status were manipulated at two separate time points. Ultimately, this experimental design consisted of a 4 (attacker's motivation – fame, money, terrorism and unknown) by 3 (attack resolution status – resolved, still at risk, and unknown) between-groups factorial design.

3.1.1 Design Overview

This experiment was conducted in November, 2013 and based on a 4 (attacker's motivations) by 3 (resolution status) between-groups factorial design. Each respondent was randomly assigned to one of twelve conditions. The four levels of attacker's motivations were fame, money, terrorism, and unknown, and the three levels of resolution status were resolved, still at risk, and unknown. The unknown conditions were included as levels in both variables as no information control conditions for comparison. The experiment was submitted to the University of Southern California's Institutional IRB and it was determined that the study qualified for Exempt, Category 2 research.

The scenario unfolds over four time periods (or scenes). During Scene 1, respondents received a credit card statement in their name from a company with which they do not have account and there were charges totaling $500. During Scene 2, respondents received a voicemail from the identity theft unit of the local police department indicating an investigation was underway and they believed the respondent's computer had been compromised by the attacker, resulting in his/her identity theft. Attacker motivation was manipulated in the content of the investigator's voicemail. He indicated that the attacker was stealing the respondent's identity to either: (1) increase his visibility and reputation within the attacker community, (2) use the compromised identity to purchase luxury items, (3) use the identity to provide financial support to a middle eastern terrorist group, or (4) was unknown (control condition).

Table 4. Scenario and Manipulations (Experiment II)

Time 1	This morning in the mail you received a credit card statement in your name from a company with which you do not have an account. As you looked over the statement, you noticed several cash advances totaling $500.
Questions	PANAS
Time 2	One week following your receipt of the suspicious credit card statement, you receive the following voice mail: "Good morning, my name is Gabriel Dawson from the Identity Theft Unit of the Police Department. Our investigation into a cyber perpetrator has led us to believe your personal computer has been compromised. We believe this individual hacked into your computer and obtained access to your email account and the cache data of your online activities. In doing so, he was able to obtain your usernames, passwords, banking information, and other personal information. Our investigation thus far shows no evidence that can confirm the perpetrator's intent. (**unknown perpetrator's intent**) / Our investigation thus far shows that the perpetrator is hacking into victims' computers to increase his visibility and reputation within the attacker community. (**fame**) / Our investigation thus far shows that the perpetrator is using the victims' identities to purchase luxury items. (**money**) / Our investigation thus far shows that the perpetrator is using the victims' identities to provide financial support to a Middle Eastern terrorist group. (**terrorism**) I plan to be in touch in the coming weeks to report on the progress of our investigation. Please be vigilant in reporting to us any suspicious mail, email, or phone call. Thank you."
Questions	PANAS, risk perception, short-term behavior
Time 3	In the days following the call from the Identity Theft Unit, you notice an increase in suspicious activity. You are receiving more spam emails, junk mails and phone calls from solicitors. More notably is your receipt of a phone call from the Department of Motor Vehicles confirming the issuance of a new driver's license you did not order. You also receive a letter in the mail from the Internal Revenue Service inquiring about your filing of duplicate income tax returns, suggesting that fraudulent returns were submitted in your name.
Questions	PANAS
Time 4	Moving ahead to several weeks following the call from the Identity Theft Unit of the Police Department, you receive yet another credit card statement in the mail from a company with which you do not have an account. This statement has a $1,500 balance. It is clear that you are continuing to experience complications as a result of your identity theft and that you are still at risk. (**still at risk**) / Moving ahead to several weeks following the call from the Identity Theft Unit of the Police Department, you recently have not received any suspicious communications or an update from the police indicating whether your identity remains at risk or not. It is unclear whether you will continue to experience complications as a result of your identity theft and if this situation has been resolved. (**resolved**) / Moving ahead to several weeks following the call from the Identity Theft Unit of the Police Department, you receive a second voicemail from Gabriel Dawson at the Police Department. He is calling to inform you that the perpetrator has been arrested and they have seized all software and electronic devices containing compromised personal data, and removed all sources online containing this information. Fortunately, you are no longer experiencing complications as a result of your identity theft and the situation is resolved. (un**resolved**)
Questions	PANAS, risk perception, long-term behavior

By Scene 3, additional evidence as to how the respondents identify was being used for identity theft was presented. Lastly during Scene 4, the resolution status of the identity theft case was reported and manipulated. Subjects either (1) received another suspicious credit card statement indicating their identity was still at risk, (2) received another call from the police indicating the attacker had been arrested and that all appropriate security measure had been take to resolve their identity theft case, or (3) received no additional information, indicating the outcome was unknown (control condition).

Following each scene, respondents were asked to evaluate their current feelings in response to the identity theft scenario. In addition, following Scene 2, respondents were asked to evaluate their perceived risk and intended short-term behavioral changes, if any. Also following Scene 4, respondents were asked to assess their perceived risk and long-term behavioral intentions. At the close of the experiment, respondents were asked to provide basic demographic information and answer questions regarding their cyber experiences and what measures they take to currently protect themselves from identity theft. A complete description of all four scenes, including the manipulations and questions following each scene, is provided in Table 4.

3.1.2 Measures

Respondents' current feelings, risk perception, intended short-term behavior and long-term behavior were measured. Details of the items in each measure are included in Table 5.

Affect. Positive Affect Negative Affect Scale (PANAS) (Watson, et al., 1988) was included following each scene to measure self-reported emotion. Only the 10-item negative affect scale was included. Each item was rated from 1(not at all) to 5 (extremely). Principal axis factoring was performed on the ten negative items of the PANAS scale from Scene 1 through Scene 4. Eight items were extracted when the number of factors was constrained to one. The two items not included in the factor were ashamed and guilty. The eight items were internally consistent with a Cronbach's alpha = .93, .92, .92 and .95, for each scene respectively.

Table 5. Measures of Experiment II

Scales	Items
Negative affect	scared, afraid, upset, distressed, jittery, nervous, ashamed, guilty, irritable, hostile
Risk perception	(1) it is just amount of time before my personal financial information is obtained
	(2) credit card fraud is very common
	(3) credit card fraud creates a major financial loss for consumers and credit card companies
	(4) identity theft is a major threat to personal privacy
	(5) identity theft cases are difficult to resolve
	(6) identity theft typically results in long-term inconveniences to the victim
	(7) the risk of identity theft is not of concern to me
	(8) if my identity is stolen, I will have to spend a lot of money fixing the problem
Short-term behavioral intentions	(1) contact the credit card company
	(2) contact the consumer credit reporting agencies
	(3) call the police
	(4) contact the Department of Motor Vehicles (DMV)
	(5) contact the Social Security Administration (SSA)
	(6) contact the Internal Revenue Service (IRS)
	(7) do nothing
	(8) cancel all your credit cards
	(9) discontinue online financial transactions
	(10) other (text box)
Long-term behavioral intentions	(1) I will use my credit card for purchases significantly less than before
	(2) I will prefer to pay for purchase items in cash
	(3) I will request the free 90 days "fraud alert" service from one of the consumer credit reporting agencies that notifies me of any request for a new line of credit in my name
	(4) I would be willing to pay $10/month ($120/year) to subscribe to a protection service that lowers my risk of identity theft
	(5) I will check my credit more often than before
	(6) I will use pseudonyms in my social network accounts
	(7) I will not visit websites with which I am not familiar
	(8) I will not make online transactions that require my personal information (e.g., online shopping, online banking, apply for credit card)
	(9) I will install better protection software on my computer
	(10) I will regularly clean and delete unnecessary documents, emails, and websites in my cache on my computer
	(11) I will use completely different password for each of my online accounts and change them regularly
	(12) I would be willing to pay for an identity theft protection service that notifies me of any requests for a new line of credit in my name

Risk Perception. An 8-item Likert scale about perceived risk of identity theft was included after scene 2 and scene 4; respondents indicated agreement on a 6-point scale from 1 (strongly disagree) to 6 (strongly agree).Factor analysis was also performed on eight items of risk perception for scene 2 and scene 4. Five items (item 3, 4, 5, 6, 8) were extracted as a factor when the number of factor was constrained to one. Cronbach's alpha = .81 and .83 for scene 2 and scene 4 respectively.

Short-term behavioral intention. Following scene 2, respondents were asked to check from ten items of actions they would intend to take if a suspicious credit card statement was received.

Long-term behavioral intention. A 12-item Likert scale about long-term intended behavior were included following scene 4; respondents indicated agreement on a 6-point scale from 1 (strongly disagree) to 6 (strongly agree).Nine (item 1, 2, 3, 4, 5, 8, 9, 11, 12) out of the twelve items were extracted as a factor when factor number was constrained to one. Cronbach's alpha = .84.

Table 6. Demographic Information and Cyber-related Experience (Experiment II)

Variables (N = 419)	Response category	Number and percentage
Have you ever had an account opened fraudulently in your name that you know of?	Yes	32 (7.6%)
	No	386 (92.1%)
Do you currently pay for an identity theft protection service (e.g. LifeLock, TrustedID, Equifax ID patrol)?	Yes	25 (6.0%)
	No	393 (93.8%)
Do you have a personal computer?	Windows	356 (85%)
	Mac	57 (13.6%)
	don't have	4 (1.0%)
Do you have a credit card?	more than one	169 (40.3%)
	only one	132 (31.5%)
	don't have	117 (27.9%)
Sex	male	233 (55.6%)
	female	103 (44.2%)
Education	less than high school	2 (.5%)
	high school	120 (28.6%)
	2-year college	89 (21.2%)
	4-year college	167 (39.9%)
	master's degree	34 (8.1%)
	PhD degree	6 (1.4%)
Personal annual gross income range before tax	below $20,000/year	131 (31.3%)
	$20,000 - $29,999/year	84 (20.0%)
	$30,000 - $39,999/year	54 (12.9%)
	$40,000 - $49,999/year	50 (11.9%)
	$50,000 - $59,999/year	30 (7.2%)
	$60,000 - $69,999/year	23 (5.5%)
	$70,000 - $79,999/year	19 (4.5%)
	$80,000 - $89,999/year	7 (1.7%)
	$90,000/year or more	20 (4.8%)
Age	range	18-114
	percentiles	25^{th} 24
		50^{th} 29
		75^{th} 39

3.1.3 Respondents

The experiment was hosted on Qualtrics.com and subjects were collected through AMT. Four hundred and twenty eight adult subjects participated in the experiment, and were compensated $0.75 for their time. Nine subjects were removed for not completing the experiment. Four hundred and nineteen subjects were included in the analysis. Table 6 presents demographic information for the sample.

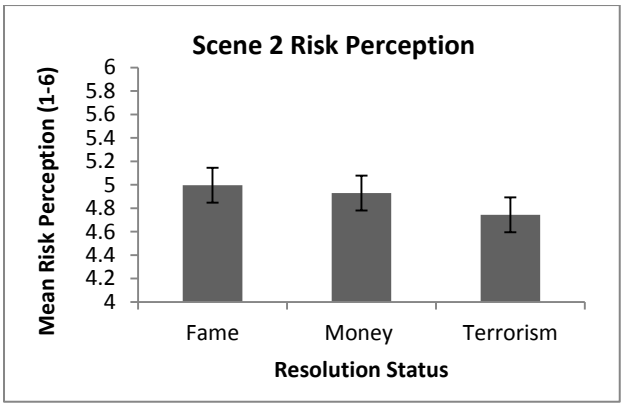
The number of respondents in each of the twelve conditions ranged from 29 to 41. Again, composite scores using equal weighting were calculated for three dependent variables: (1) negative affect, (2) risk perception, (3) long-term behavioral intention; scores for short-term behavior were calculated by counting the number of actions respondents checked from the eight items presented.

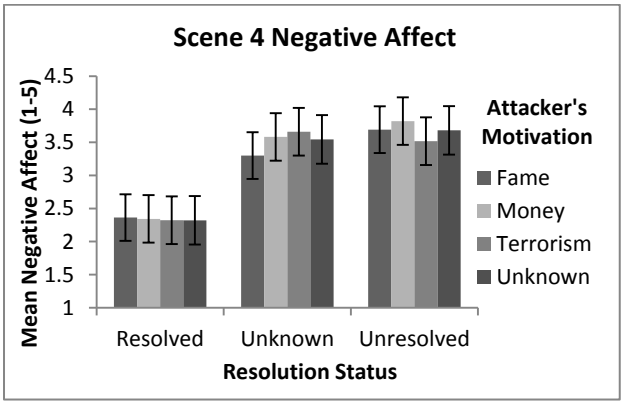
3.2 Results

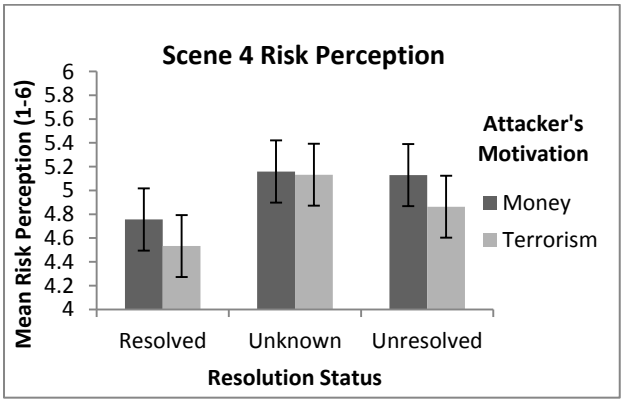
OLS regression analyses were conducted to predict the four dependent variables (affect, risk perception, short-term behavioral intentions, and long-term behavioral intention) from the two manipulated variables (the attacker's motivations---fame, money, terrorism or unknown and resolution status --- resolved, still at risk, or unknown), and respondents' sex. To examine the influence of the attacker's motivations, three orthogonal contrasts were created and entered the regressions as independent variables: (1) unknown vs. fame, money and terrorism, (2) terrorism vs. fame and money, (3) fame vs. money. To examine the influence of resolution status, the resolved condition was contrasted against the unresolved and unknown conditions.

Results indicate that following Scene 2 respondents perceived the risk of identity theft to be lower when the attacker's motivation was to fund terrorism compared to gaining money or fame (standardized $\beta = .109$, $t = 2.307$, $p = .022$, $R^2 = .084$). No significant difference was found between the motivations funding terrorism, personal financial gain, or fame for reported negative affect and short-term behavior following Scene 2. In addition, negative affect, perceived risk and short-term behavior were not significantly different between unknown vs. fame, money and terrorism and fame vs. money. Following Scene 4, respondents reported less negative affect at Scene 4 when the identity theft case was reported as resolved compared to unresolved or uncertain (standardized $\beta = .496$, $t = 11.463$, $p < .001$, $R^2 = .262$). It was also found that the perceived risk of identity theft was lower when the outcome of the scenario was reported as resolved compared to unresolved or uncertain (standardized $\beta = .104$, $t = 2.175$, $p = .030$, $R^2 = .076$). Following Scene 4, respondents were more willing to pursue long-term behavior change, such as discontinuing online transactions that require personal information or purchasing an identity theft protection service, when the outcome of the identity theft case was unresolved or uncertain compared to the resolved condition (standardized $\beta = .098$, $t = 1.984$, $p = .048$, $R^2 = .025$). Figure 2 displays the mean negative affect, perceived risk, and long-term behavioral intentions for different attacker motivations and the scenario resolution status following Scenes 2 and 4.

Lastly, results from the regression analyses indicate that sex significantly predicts perceived risk (standardized $\beta = .256$, $t = 5.370$, $p < .001$) and short-term behavioral intentions (standardized $\beta = .135$, $t = 2.714$, $p = .007$) following Scene 2, and negative affect (standardized $\beta = .124$, $t = 2.834$, $p = .005$), perceived risk (standardized $\beta = .238$, $t = 4.959$, $p < .001$), , and long-term behavioral intentions (standardized $\beta = .121$, $t = 2.435$, $p = .015$) following Scene 4. Overall, female respondents reported higher negative affect, more perceived risk, and a greater intention to seek help (short-term) and pursue online identity protection (long-term).







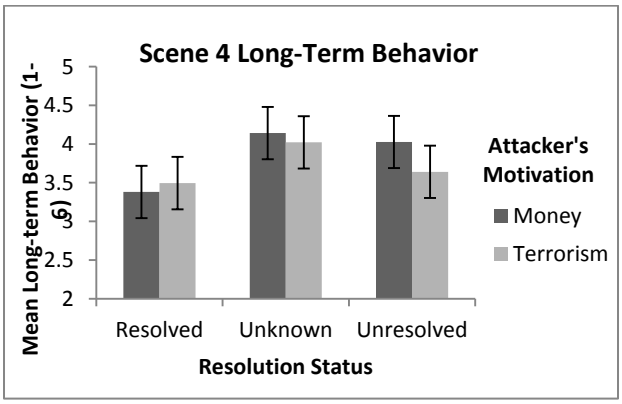


Figure 2. Mean risk perception, negative affect, and long-term behavior for attacker's motivation and resolution status. Note: Error bars are +/- 2 SE.

3.3 Discussion

Responses to a cyber-based identity theft attack in Experiment II were found to be significantly predicted by the attacker's motivations and the resolution status of the scenario. Consistent with our hypothesis, the closer and more personal the attacker's motivation was perceived to be, the greater the perceived risk of identity theft. In particular, respondents who were told the attacker's motivation was for personal financial gain interpreted the scenario as more realistic and familiar compared to the attacker who stole the respondent's identity to fund terrorism.

No difference in response was found across respondents who were told the attacker's motivations were for money, fame, or unknown. We suspect this might be a result of all three motivation types being driven by the same underlying means – money. That is, the theft of money is necessary to meet the desired end, whether the attack motivation is for personal gain, fame or an unknown reason. Furthermore, these three factors are perceived to be more personally motivated compared to the politically-driven motivation to fund terrorism (Brenner, 2007).

Following Scene 4 when scenario resolution status was manipulated, the findings suggest that lower levels of negative affect and perceived risk resulted from the resolved scenario compared to the unresolved and unknown scenarios. Consistent with our hypothesis, it is reasonable to assume that resolved outcomes create feelings of security for respondents and are less likely to induce any desire or need for behavioral change. This was reflected in responses by those in the resolved condition who following Scene 4 perceived the risk of identity theft to be lower as well as were less inclined to make behavioral changes that would protect their identity for the long-term.

We also found that respondents responded similarly to the unresolved and unknown outcome conditions. It is reasonable to expect that the level of uncertainty associated with the unresolved and unknown outcomes is perceived similarly, and for this reason respondents are more willing to engage in long-term behavioral change. Interestingly though, respondents indicated that on average they were only "somewhat willing to agree" to engage in long-term behavioral changes. This is consistent with recent poll results showing that the majority of U.S. adults (93 %) recognize identity theft is a growing problem, yet are failing to practice simple safeguards; e.g. more than half (55 %) of respondents indicated that they do not always check to see if a website is secure before shopping online, and more than three out of five respondents who had online accounts (63 %) do not use a unique password for each of their online accounts (PRNewswire, 2013).

Lastly, as in Experiment I, we found an anticipated sex effect, indicating that female respondents reported greater negative affect, greater perceived risk, greater intent to pursue short-term behavior and long-term behavior. This finding continues to be consistent with results showing that males have a greater tendency to engage in risky behaviors online (Milne et al., 2009) and females tend to demonstrate higher security procedure compliance (Herath and Rao, 2009).

Victims of identity theft are often in the position where they must take the initiative to address and manage the privacy breach. Services are available to support their needs, such as the police department referenced in the scenario, yet the process of resolution is largely self-motivated. As such, the policy implications of our findings provide potential insight into the type

of response to expect from identity theft victims, and how to communicate with such victims given awareness of the attacker motivation and attack resolution. More specifically, victims of attacks for which the outcome is uncertain and the attacker is motivated by financial gain are more likely to modify their behavior and seek out the support of identity theft-related social services. To the contrary, victims for which the attacker motivation is more distant (not financially driven) and the attack outcome is resolved, additional effort by the social service providers, and in turn additional money, will likely be needed to generate the desired behavioral response for managing the ongoing risks of identity theft. Again, findings suggest that females have the potential to be more inclined to support policy recommendations, yet additional research is needed to fully understand the moderating effects of demographic variables.

4. CONCLUSIONS

These experiments were designed to explore how individual computer users' responses to common cyber-based financial fraud and identity theft scenarios are influenced by attacker characteristics and attack mode (Experiment I) and attacker motivation and attack resolution status (Experiment II). The same response constructs were used in both studies, but were defined slightly differently given variations in the scenario contexts.

Both of these experiments utilized a scenario simulation methodology and an experimental manipulation design with concrete, realistic stimulus materials to explore respondents' predictions about their feelings, perceived risk and behavioral intentions to respond to the simulated financial fraud and identity theft attacks. As suggested by construal theory (Trope & Liberman, 2003, 2010; Trope, Liberman, & Wakslak, 2007), it is hard for people to assess their reactions when the context is more distant and unobservable. While surveys and focus groups are useful, one of their limitations is the reliance on cognition in the absence of any attention to affect (Slovic et al., 1994). The scenario simulation methodology is designed to present scenarios that are both believable and effective in evoking emotional responses from respondents.

Across the two experiments, results indicate that attacker characteristics and attack mode (Experiment I) and attacker motivations (Experiment II), influenced the perception of vulnerability of respondents to the financial fraud and identity theft scenarios. In Experiment I, the pictorial identification of the attacker resulted in more proximal and concrete interpretations of the attacker characteristics, resulting in lower negative affect. In Experiment II the more concrete and "real" attacker motivations were associated with higher perceived risk. Interestingly, the use of pictures in the characterization of the manipulated variables changed the direction of the reaction to the cyber attack.

Studies of cyber security have shown that management of affective reactions and perceived risk strongly influence individual users' decisions. For example, individual users experiencing lower perceived risk were more likely to purchase a product online; likewise, users feeling greater negative affect were less likely to sign-up for online banking services (Kim, Ferrin & Rao, 2008; Lee, 2009). This result is consistent with our experimental findings, suggesting that when respondents felt that they were vulnerable, they responded with heightened behavioral response. More specifically, in Experiment I all respondents felt some level of vulnerability in response to the cyber-based bank hacking scenario and for this reason had an expectation that the bank would take action to mitigate the consequences of the attack. The respondents' behavioral intentions following the attack varied as a function of the manipulated characteristics of the simulated financial fraud attack scenario. Similarly, in Experiment II, respondents recognized the perceived risk associated with identity theft and expressed a willingness to engage in long-term behavior change. Again, the degree of intended behavior change varied relative to the resolution status indicated in the simulated identity theft scenario.

There is limited research on the influence of attacker attributes on individual user decision making. The scenario simulation approach used in our experiments presents a more emotionally evocative and realistic method for assessing individual reactions and decision making compared to traditional descriptive survey studies and post-hoc field studies. However, given the global reliance and dependence on the internet and the frequency with which cyber attacks occur, a study of actual victims emotional reactions, perceived risk and decision making following a real attack would be an important next research step.

In addition, more studies are needed to further understand whether the identified relationships are generalizable to other cyber threat scenarios. Given that safety and security in the cyber context are abstract concepts, it would be worthwhile to further explore how attacker attributes influence reactions and decision making in response to cyber attacks. This research design also could be used to evaluate differences across a varied set of cyber-based attacks to examine the robustness of the relationships identified in our research. Lastly, future studies could also be designed to specifically address policies designed to assess privacy preferences given attacker attributes. This experimental design would be more directed at studying specific policy tools and educational approaches for addressing the cyber threat in the present, similar to work reported 15 years ago by Ackerman, Cranor and Reagle (1999).

5. ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation under grant number SES-1314644. It was also supported by the U. S. Department of Homeland Security through the National Center for Risk and Economic Analysis of Terrorism Events under the cooperative agreement number 2010-ST-061-RE0001. However, any opinions, findings, conclusions, and recommendations in this document are those of the author and do not necessarily reflect views of the National Science Foundation or the U. S. Department of Homeland Security. We would like to thank Lauren Ladd-Reinfrank for her support in the planning and execution of Experiment I.

6. REFERENCES

[1] Ackerman, M.S., Cranor, L.F., & Reagle, J. (1999). Privacy in e-commerce: examining user scenarios and privacy preferences. *EC '99 Proceedings of the 1st ACM Conference on Electronic Commerce,* 1-8.

[2] Bhatnagar, A., Misra, S., & Rao, H. R. (2000). On risk, convenience, and Internet shopping behavior. *Communications of the ACM, 43*(11), 98-105.

[3] Bourque, L. B., Regan, R., Kelley, M. M., Wood, M. M., Kano, M., & Mileti, D. S. (2013). An examination of the effect of perceived risk on preparedness behavior. *Environment and Behavior, 45*(5), 615-649.

[4] Brenner, S. W. (2007). " At Light Speed": Attribution and Response to Cybercrime/Terrorism/Warfare. *The Journal of Criminal Law and Criminology*, 379-475.

[5] Cameron, L., & Shah, M. (2013). Risk-taking behavior in the wake of natural disasters. *National Bureau of Economic Research*.

[6] D'Amico, A., Whitley, K., Tesone, D., O'Brien, B., & Roth, E. (2005). Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts. *Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.*

[7] Fischhoff, B., Gonzalez, R. M., Small, D. A., & Lerner, J. S. (2003). Judged terror risk and proximity to the World Trade Center *The Risks of Terrorism* (pp. 39-53): Springer.

[8] Garbarino, E., & Strahilevitz, M. (2004). Gender differences in the perceived risk of buying online and the effects of receiving a site recommendation. *Journal of Business Research, 57*(7), 768-775.

[9] Hale, C. (1996). Fear of crime: A review of the literature. *International Review of Victimology, 4*(2), 79-150.

[10] Henson, B., Reyns, B. W., & Fisher, B. S. (2013). Fear of Crime Online? Examining the Effect of Risk, Previous Victimization, and Exposure on Fear of Online Interpersonal Victimization. *Journal of Contemporary Criminal Justice, 29*(4), 475-497.

[11] Herath, T., & Rao, H. R. (2009). Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. *Decision Support Systems, 47*(2), 154-165.

[12] Ho, M. C., Shaw, D., Lin, S., & Chiu, Y. C. (2008). How do disaster characteristics influence risk perception? *Risk Analysis, 28*(3), 635-643.

[13] Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems, 44*(2), 544-564.

[14] Kung, Y. W., & Chen, S. H. (2012). Perception of earthquake risk in Taiwan: Effects of gender and past earthquake experience. *Risk Analysis, 32*(9), 1535-1546.

[15] Lee, M.-C. (2009). Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit. *Electronic Commerce Research and Applications, 8*(3), 130-141.

[16] Leiserowitz, A. (2006). Climate change risk perception and policy preferences: the role of affect, imagery, and values. *Climatic Change, 77*(1-2), 45-72.

[17] Liu, L., Yu, E., & Mylopoulos, J. (2003). Security and privacy requirements analysis within a social setting. *Paper presented at the Requirements Engineering Conference, 2003. Proceedings. 11th IEEE International.*

[18] Milne, G. R., Labrecque, L. I., & Cromer, C. (2009). Toward an understanding of the online consumer's risky behavior and protection practices. *Journal of Consumer Affairs, 43*(3), 449-473.

[19] Norris, F. H., Stevens, S. P., Pfefferbaum, B., Wyche, K. F., & Pfefferbaum, R. L. (2008). Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *American Journal of Community Psychology, 41*(1-2), 127-150.

[20] Nykodym, N., Taylor, R., & Vilela, J. (2005). Criminal profiling and insider cyber crime. *Digital Investigation, 2*(4), 261-267.

[21] Peters, E., & Slovic, P. (1996). The role of affect and worldviews as orienting dispositions in the perception and acceptance of nuclear Power1. *Journal of Applied Social Psychology, 26*(16), 1427-1453.

[22] PRNewswire. (2013). Many consumers fear identity theft yet still engage in risky behavior. Available online at http://www.prnewswire.com/news-releases/many-consumers-fear-identity-theft-yet-still-engage-in-risky-behavior-228595151.html.

[23] Reponses, P., Model, A. N., & Westbrook, L. (2012). Private crises/public reponses: A nascent model. *Proceedings of the American Society for Information Science and Technology, 49*(1), 1-12.

[24] Rosoff, H., Cui, J., & John, R. S. (2013). Heuristics and biases in cyber security dilemmas. *Environment Systems and Decisions, 33*(4), 517-529.

[25] Rosoff, H., Siko, R., John, R., & Burns, W. J. (2013). Should I stay or should I go? An experimental study of health and economic government policies following a severe biological agent release. *Environment Systems & Decisions*, 1-17.

[26] Sjøberg, S., & Schreiner, C. (2005). *Young people and science.* Paper presented at the Attitudes, values and priorities. Evidence from the ROSE project. Keynote presentation at EU's Science and Society Forum.

[27] Skogan, W. G., & Maxfield, M. G. (1981). *Coping with crime: Individual and neighborhood reactions*: Sage Publications Beverly Hills, CA.

[28] Slovic, P. (1987). Perception of risk. *Science, 236*(4799), 280-285.

[29] Slovic, P., Fischhoff, B., Lichtenstein, S., & MacGregor, D. (2004). Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality. *Risk Analysis, 24* (2), 311-322.

[30] Slovic, P., Fischhoff, B., & Lichtenstein, S. (1980). Facts and fears: Understanding perceived risk. *Societal Risk Assessment* (pp. 181-216): Springer.

[31] Slovic, P., Fischhoff, B., & Lichtenstein, S. (1984). Behavioral decision theory perspectives on risk and safety. *Acta Psychologica, 56*(1), 183-203.

[32] Terpstra, T. (2011). Emotions, trust, and perceived risk: Affective and cognitive routes to flood preparedness behavior. *Risk Analysis, 31*(10), 1658-1675.

[33] Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110*(3), 403.

[34] Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440.

[35] Trope, Y., Liberman, N., & Wakslak, C. (2007). Construal levels and psychological distance: Effects on representation, prediction, evaluation, and behavior. *Journal of Consumer*

Psychology: The Official Journal of the Society for Consumer Psychology, 17(2), 83.

[36] Vlek, C., & Stallen, P.-J. (1980). Rational and personal aspects of risk. *Acta Psychologica, 45*(1), 273-300.

[37] Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063.

Towards Continuous and Passive Authentication via Touch Biometrics: An Experimental Study on Smartphones

Hui Xu* Yangfan Zhou*† Michael R. Lyu*‡
*Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications,
Shenzhen Research Institute, The Chinese University of Hong Kong
†MoE Key Laboratory of High Confidence Software Technologies (CUHK Sub-Lab)
‡Dept. of Computer Science & Engineering, The Chinese University of Hong Kong

ABSTRACT

Current smartphones generally cannot continuously authenticate users during runtime. This poses severe security and privacy threats: A malicious user can manipulate the phone if bypassing the screen lock. To solve this problem, our work adopts a continuous and passive authentication mechanism based on a user's touch operations on the touchscreen. Such a mechanism is suitable for smartphones, as it requires no extra hardware or intrusive user interface. We study how to model multiple types of touch data and perform continuous authentication accordingly. As a first attempt, we also investigate the fundamentals of touch operations as biometrics by justifying their distinctiveness and permanence. A one-month experiment is conducted involving over 30 users. Our experiment results verify that touch biometrics can serve as a promising method for continuous and passive authentication.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; D.4.6 [**Software**]: Security and Protection

General Terms

Human Factors, Security, Experimentation

Keywords

Smartphone, Continuous Authentication, Touch Biometrics

1. INTRODUCTION

Smartphones are becoming more and more popular in people's daily life. According to a recent report [31], the number of smartphone users has reached 56% of the American adult population, and smartphone sales continue to grow radically [11]. As a result of the extensive usage of smartphones, much of our sensitive and private information is kept by our phones. This inevitably poses great security risks to smartphone users [8, 13, 35].

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

To mitigate the risk of malicious user access, most smartphone systems adopt a traditional access control mechanism: Before using a phone, a user needs to unlock its screen with a password or a lock pattern (*i.e.*, several dots in the screen that should be visited in sequence in one finger move). Since a user may use her phone quite often in her daily life, password or lock pattern should be designed simple enough to facilitate the frequent unlock operations. This severely degrades the strength of the access control mechanism. Malicious users can break into the phone simply via peeping [9], or the smudge attack [5].

An enhanced mechanism, namely *continuous authentication* [14, 27], can be more effective in combatting malicious user access. It keeps authenticating the current user during system runtime, thus greatly increasing the complexity of potential intrusions. Examples for such mechanism include requiring fingerprint[1] or face authentication frequently, asking for the answers of a set of pre-defined security problems or passwords, or connecting to an accessory device owned by the valid user[2]. However, these approaches are either too intrusive (*e.g.*, keep asking for password or fingerprint) or costly (*e.g.*, require an extra device like fingerprint sensor or the "Skip"), not to mention the extra energy required to drive the sensors.

We observe that the user operations on touchscreen can be utilized for continuous authentication, with no requirement for extra hardware or user attention. As the dominant human-to-smartphone interface [34], touchscreen is equipped on most smartphones. Moreover, modern touchscreens can produce rich data to describe how users touch, including the curve, the timing, the size and the pressure of a touch operation. Such data can be collected in the background and analyzed to discriminate different users. In other words, while the user performs her normal operations, the authentication proceeds continuously without her notice, *i.e.,* in a passive way.

Using touch operations for continuous authentication has been suggested recently in [14], where a single type of operations (strokes or slides) is considered. Some promising results have been reported. For example, a 13% equal error rate (EER) for one single stroke, and 2% to 3% for 11 consequent strokes can be achieved [14]. However, stroke is not the only type of touch operations. They can also include other types, such as pinch and handwriting. Hence, consid-

[1]Note that recently Apple and Samsung have embedded fingerprint sensor into their smartphones.
[2]For example, the "Skip" device introduced by Motorola for MotoX phone.

ering only strokes is not enough to continuously authenticate the user as she can perform other types of operations. A seamless continuous authentication mechanism should take multiple types of operations into account. Moreover, previous investigations (*e.g.*, [9],[14]) have based their designs on a rather straightforward idea that touch operations can be employed to identify users. Yet, the biometric properties of touch operations have not been comprehensively evaluated.

Our work, in contrast, takes advantage of multiple types of touch data to model a user. As a first attempt, we further investigate the underlying fundamentals of touch operations as biometrics by justifying their two critical properties: *distinctiveness* and *permanence*. In other words, we evaluate whether the data features are distinctive among various users, and whether the data features collected from the same user are temporally stable. Both properties are prerequisites for biometrics [17].

To this end, we have conducted a real-world experiment involving over 30 users for one month. Our results confirm that it is promising to implement a continuous authentication mechanism based only on the touch data collected during normal user operations.

The contributions of this paper are as follows:

- This work serves as the first attempt to comprehensively evaluate the biometric properties of touch data, and we study how such data can be used for continuous authentication.
- We propose a set of methods to model the multiple types of touch data via a separation-of-concern solution, which is quite effective.
- The findings and data from our real-world experiment involving over 30 users are publicly available, which can facilitate further follow-up work.

The rest of the paper is organized as follows. Section 2 provides the adversary model and some preliminaries of touch biometrics. Section 3 overviews the framework of touch-based authentication and goes through details about the feature extraction and classification method. In Section 4, we evaluate the performance of touch biometrics in distinctiveness, permanence and authentication error rate based on the framework. The related work is discussed in Section 5. Section 6 concludes our research and suggests potential future work.

2. BACKGROUND

In this section, we briefly introduce the adversary model and some technical preliminaries including smartphone touch operations, biometrics, and performance metrics.

2.1 Adversary Model and Assumptions

In this paper, we assume the following adversary. A malicious attacker has gained access to a person's smartphone equipped with a touchscreen. The smartphone is either unprotected (*e.g.,* no PIN) or the attacker has got into possession of the authentication secret, for instance by shoulder surfing the owner. The attacker can then perform undesirable actions with the device violating the owner's privacy (*e.g.,* browsing photos, reading SMS or e-mails). Afterwards, the phone's screen can be turned off and put back to its original place, appearing as if it was never touched. The owner will have no chance to figure out that it has been used by someone else. In this way, the owner's privacy could be severely violated. Our work targets such situations and tries to make this kind of manipulation impossible by analyzing touch behavior.

Table 1: Example of raw event data collected when tapping "1" and "2" on soft keyboard

Tap	Time	Position		Size	Pressure
		X	Y		
1	122382	62.869	550.312	0.169	0.233
1	122444	67.892	553.328	0.169	0.2
1	122461	70.057	550.008	0.067	0.067
1	122503	70.057	550.008	0.067	0.067
2	122731	202.578	553.308	0.141	0.167
2	122794	204.591	556.305	0.141	0.2
2	122811	204.574	554.170	0.141	0.2

2.2 Touch Operations

The smartphone systems accept user commands through interpreting touch. According to our knowledge, the most frequently used operations include *keystroke*, *slide*, *pinch*, and *handwriting*.

- *Keystroke:* A keystroke is a finger tap on the screen. Typical scenarios include using soft keyboard and unlocking screen with PIN.
- *Slide:* A slide is a finger move on the screen. A lot of applications use slide for navigating documents, *e.g.*, web pages, photo albums, messages, and contact list.
- *Pinch:* A pinch is a two-finger gesture typically used for zooming functionality.
- *Handwriting:* Handwriting is an important alternative input method on smartphone to enter characters.

When a touch operation is performed, the smartphone hardware automatically generates a set of data and reports them to the operating system as *raw events*. Taking Android as an example, a raw event reports the data of the position, pressure, and size of a touch, as well as a timestamp. The operating system generally extracts touch operations intended by the user by interpreting such raw events. Each row in Table 1 shows the data of a raw event. We observe in our practice that the *time* and *position* data are fine-grained, while the *size* and *pressure* are coarse-grained. To avoid noise, we choose to use statistical information (*e.g.*, average or standard deviation) of the size and pressure data instead of subtle data changes in the feature extraction process.

In practice, one single touch operation generates a series of raw events. Their positions form a trajectory sequence. We call the sequence of the corresponding raw event data a *touch data sequence* of the touch operation. Touchscreen can produce raw events every few milliseconds when being touched. As a result, even the simplest touch operation can generate quite a few raw events. Table 1 shows an example of raw events collected when tapping "1" and "2" on the soft keyboard. In this example, the tap on "1" and "2" produce four and three raw events. We will discuss how we model a touch operation based on the touch data sequence it generates in Section 3.1.

2.3 Biometrics

Biometrics refers to the automatic recognition of individuals based on their physiological and/or behavioral characteristics [17]. Common types of biometrics include face, fingerprint, hand geometry, iris, keystroke, signature, and voice [16]. When a biological characteristic qualifies to be a form of biometrics, it should generally bear the following four properties [17].

- *Universality*: Every person has the characteristic.
- *Distinctiveness*: Any two persons are distinguishable in terms of the characteristic.
- *Permanence*: The characteristic is stable over a period of time.
- *Collectability*: The characteristic can be measured in numbers.

Touch operation can be considered as of behavioral biometrics. Its universality and collectability are obvious, while its distinctiveness and permanence need to be assessed, which is a major focus of our work.

Note that there are also other issues that need to be considered for a practical biometric system, for example, recognition speed, overhead, and user-friendliness [17]. These implementation considerations are not the focus of this work.

2.4 Performance Metrics

Accuracy and error rate are two straightforward metrics for authentication performance. However, their information is rather limited and must be interpreted with much caution. It is therefore necessary to introduce the concepts of false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER) and receiver operating characteristic (ROC), which are more meaningful [24]. These terms are defined as follows:

- *FAR*: The rate that an attacker is wrongly accepted as the valid user.
- *FRR*: The rate that the valid user is wrongly rejected as an attacker.
- *EER*: The rate at which FAR and FRR are equal. In practice, FAR and FRR are sensitive to system settings and correlated with each other. FAR will usually increase as FRR decreases, and *vice versa*. EER is a metric of the trade-off between of FAR and FRR, which is widely used for indicating the performance of real authentication systems.
- *ROC*: A graphical plot that visualizes the performance of a binary classifier as its discrimination threshold varies. ROC is created by plotting the fraction of the true positive rate (*i.e.*, rejection rate when the user is invalid) vs the false positive rate (*i.e.*, rejection rate when the user is valid), at various threshold settings [1]. ROC is a more complicated indicator, which reflects the performance of a system under different settings.

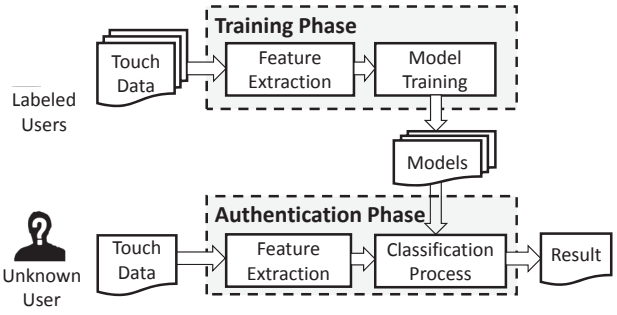


Figure 1: Overview of touch-based authentication approach

3. TOUCH DATA-BASED USER AUTHENTICATION

Our idea of using touch data for continuous authentication includes two phases: the training phase and the authentication phase. In the training phase, a number of *labeled* touch data (*i.e.*, the data together with whether it comes from a valid user) are processed so as to model the valid user. In the authentication phase, the touch data, which may come from the valid user or an attacker, are labeled according to the models generated in the training phase. In this way, we can authenticate the corresponding user of the touch data. Fig. 1 overviews the touch-based user authentication approach.

Centric to this approach is a statistical pattern recognition procedure that can discriminate different users according to the touch data. To design an effective touch data-based user authentication approach, two key steps need to be addressed: 1) how to model the user characteristics from the touch data, *i.e.*, what kind of features should be extracted from the data. 2) how to recognize users according to these features. We discuss these two issues in what follows.

3.1 Feature Extraction

Touchscreen can catch every subtle user touch and generate corresponding touch data sequence. We may directly consider touch data sequence as the basic granularity and model the user accordingly. However, since different sequences may belong to different types of touch operations, they may contain quite different characteristics. For example, a slide operation with one finger move is quite different from a pinch operation with two fingers. In order to address this problem, we propose a separation-of-concerns approach which considers each type of touch operations separately. In this way, each type of touch operations can be modeled separately with its corresponding sequence of raw events.

Let X denote the data of a raw event, where $X = [\text{Time}, \text{Position}_x, \text{Position}_y, \text{Pressure}, \text{Size}]$. Let $\{X_1, X_2, ..., X_n\}$ denote a sequence of raw events that jointly form a touch operation. Let $F = [\text{feature}_1, \text{feature}_2, ..., \text{feature}_m]$ denote the feature vector of a touch operation. We should find how to map $\{X_1, X_2, ..., X_n\}$ to F, so that F can well describe the characteristics of the touch operation. In what follows, we will discuss the design of such a mapping according to the specifics of each type of touch operations.

3.1.1 Features of Keystroke

Keystroke operation typically involves a series of taps on the soft, on-screen keyboard. Keystroke dynamics on hard-

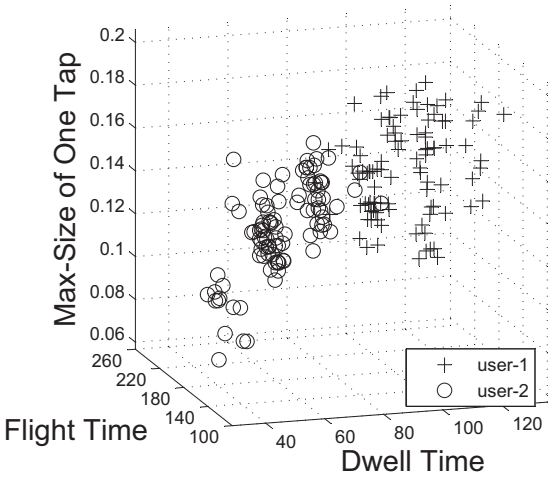


Figure 2: Keystroke feature vectors of 2 users in 3-dimensional space when tapping "1" within a number sequence "123456"

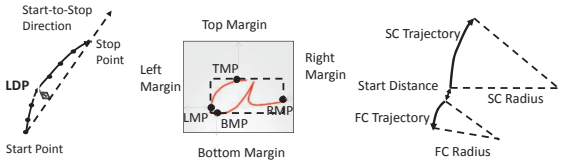
ware keyboard is a type of biometrics well studied in the literature [4, 22], which sheds light to our study on software keyboard. We adopt two features proven effective in the hardware keystroke dynamic field: the *dwell time* and *flight time* features. The former considers the duration of a keystroke and the latter considers the time interval between successive keystrokes. Even though some new features specially tailored for touchscreen based keystrokes have been proposed (*e.g.*, the detailed touch locations of each key [10]), there is no enough evidence to show that the recognition accuracy can be improved considerably [10]. Hence, we don't include these new features in our model.

The upper-left corner of Table 2 shows the four typical features for keystroke operation we propose. Besides dwell time and flight time, the other two features are self-explained by their names. As a demonstrating example, Fig. 2 shows the feature vectors extracted from 2 different users when they perform keystroke operations. We can easily see that different people have quite different characteristics in terms of the features we propose.

3.1.2 Features of Slide

A slide operation is a finger move from a start point to a stop point on the screen, *i.e.*, a curve. Besides these two points, we also consider the largest deviation point (LDP) in the slide curve. An LDP is the point that is farthest to the straight line between the start point and the stop point of the slide curve. Fig. 3(a) shows an example of such an LDP. The LDP can, to some extent, describe the curvature of the slide. Hence, we choose to extract features based on these three points. Our extraction process is designed as follows.

First, we consider the positions of these three points, and thus introduce the *trajectory features*. Trajectory features are the features that reflect the directional information of finger moving and those that measure the length of the moving trajectory. The latter is measured by the sum of the line segments between every two consecutive raw events occurring during the finger move. Secondly, we consider the dynamics of the slide move along these three points. Specially, we consider the pressure, size and velocity along them.



(a) Slide (b) Handwriting (c) Pinch

Figure 3: Demonstration of key metrics during feature extraction

Thirdly, there are several statistical features that have been taken into account. For example, the standard deviation of touch pressure occurring during a slide can reflect the distribution of touch strength. Table 2 provides the 37 suggested features for slide.

3.1.3 Features of Handwriting

Input via writing on the screen is an important input method for smartphones. Naturally, how to model such operations is the area of handwriting forensic. Handwriting forensic identifies handwriting through the analysis of various aspects of writing, including the arrangement, slant, baseline alignment, design of alphabets [32]. In this work, we also extract handwriting features with the handwriting forensic approach. We omit those features that are not computationally available [32] and customize 42 features for handwriting authentication, as provided in Table 2. Specifically, we consider the leftmost, rightmost, topmost, and bottommost points of a handwritten letter (denoted by LMP, RMP, TMP, and BMP, respectively). Fig. 3(b) demonstrates these four points of a handwritten "*a*".

Similar to slide operation, we propose the trajectory features of these four points, as well as dynamics of the finger move along these points. We also consider the statistical features of raw events which occur during the handwriting.

3.1.4 Features of Pinch

The trajectory of a pinch operation includes two curves, since it involves two fingers. The features of a pinch naturally include the features of both curves. The features of each curve can be extracted similarly as a slide. We also consider the features that can describe the correlation between the curves, as they are generated by two fingers of the same user. For example, we consider *start distance* and *stop distance*, which are the distances between two fingers when the pinch starts and stops respectively.

We notice some people would pinch with thumb and index finger, while others with index finger and middle finger, which will cause quite different characteristics of the resulting curves. Instead of distinguishing the two curves with finger name, we distinguish the two curves by their positional information: The curve with the start position on the left-hand side to the start position of the other curve is named the first curve (FC), and the other curve is named the second curve (SC). There are in total 49 features we propose for modeling the pinch as listed in Table 2.

In the discussions above, we have provided a set of features for each type of touch operations based on their specifics. It is worth noting that these features may not all be effective for user authentication. In our experimental study, we will evaluate these features and select a subset for modeling each type of touch operations.

Table 2: The features we proposed for touch operations(Pos. and Traj. stand for position and trajectory, respectively). For each feature, we present the feature evaluation result in accuracy according to Section 4.2.

Keystroke Features			Handwriting Features			Pinch Features		
Feature Name	Accurary (%)	Ranking	Feature Name	Accurary (%)	Ranking	Feature Name	Accurary (%)	Ranking
Max-Size of One Tap	18.3761	1	Left Margin	12.3	27	FC Start Point Pos. X	19.2	3
Max-Pressure of One Tap	9.9343	4	Right Margin	11.1	31	FC Start Point Pos. Y	16	14
Dwell Time	13.1823	3	Top Margin	16.6	15	FC Start Point Size	14.3	20
Flight Time	13.4075	2	Bottom Margin	20.2	4	FC Start Point Pressure	13.8	23
Slide Features			LMP Size	19.8	5	SC Start Point Pos. X	13.9	22
Start Point Pos. X	20.6	1	RMP Size	9.9	35	SC Start Point Pos. Y	15.5	18
Start Point Pos. Y	16.4	7	TMP Size	23.3	1	SC Start Point Size	18	8
Start Point Size	18.7	2	BMP Size	17	12	SC Start Point Pressure	11	29
Start Point Pressure	10.1	18	LMP Pressure	8.7	36	FC Stop Point Pos. X	18.4	4
Start Point Velocity	10.6	16	RMP Pressure	4.7	38	FC Stop Point Pos. Y	14.3	21
LDP Pos. X	12.4	14	TMP Pressure	11.9	28	FC Stop Point Size	9	38
LDP Pos. Y	11.5	15	BMP Pressure	7.9	37	FC Stop Point Pressure	6.5	43
LDP Size	18.5	3	Vertical Direction	2.4	41	SC Stop Point Pos. X	16	15
LDP Pressure	10.4	17	Horizontal Direction	2.4	40	SC Stop Point Pos. Y	24.5	1
LDP Velocity	14.2	11	Avg. Size	18.2	9	SC Stop Point Size	12.6	25
Stop Point Pos. X	16.2	8	Avg. Pressure	20.9	2	SC Stop Point Pressure	9.4	37
Stop Point Pos. Y	14.5	10	Start Point Pos. X	11.5	29	FC Start Point Velocity	10.2	32
Stop Point Size	7.7	28	Start Point Pos. Y	16.6	14	FC Stop Point Velocity	8.1	41
Stop Point Pressure	5.5	30	Start Direction	3.2	39	SC Stop Point Velocity	9	39
Stop Point Velocity	8.5	26	Stop Point Pos. X	12.3	26	SC Start Point Velocity	9.8	34
Avg. Velocity	16.8	5	Stop Point Pos. Y	19	6	FC Traj. Length	15.6	16
Start-to-LDP Latency	8.7	25	Stop Direction	2	42	SC Traj. Length	16.7	13
Straight Start-to-LDP Length	9.4	21	Start-to-LMP Latency	11.5	30	FC Interval	19.6	2
Start-to-LDP Direction	4.7	32	Start-to-RMP Latency	19	7	SC Interval	18.3	6
Start-to-Stop Latency	10	19	Start-to-TMP Latency	13.8	21	FC Traj. Velocity	9.8	35
Straight Start-to-Stop Length	9.1	22	Start-to-BMP Latency	17	13	SC Traj. Velocity	11.8	27
Start-to-Stop Direction	3.4	36	Start-to-LMP Traj. Length	13	24	Start Distance	13.5	24
LDP-to-Stop Latency	14.1	12	Start-to-RMP Traj. Length	18.2	8	Stop Distance	15.6	17
Straight LDP-to-Stop Length	16.5	6	Start-to-TMP Traj. Length	16.2	16	Start Interval	8.2	40
LDP-to-Stop Direction	4	35	Start-to-BMP Traj. Length	17.8	11	Stop Interval	11	30
Straight LDP Length Ratio	7.7	27	Start-to-LMP Velocity	10.7	33	Mutual Interval	18.4	5
Start Direction	2.7	37	Start-to-RMP Velocity	13.4	23	Traj. Length Ratio	16.8	10
Stop Direction	4.2	33	Start-to-TMP Velocity	16.2	17	FC Moving Direction	3.3	46
Rotation	4	34	Start-to-BMP Velocity	14.6	20	SC Moving Direction	2.4	48
Traj. Length	17.1	4	Total Traj. Length	20.6	3	FC Moving Rotation	5.3	44
Straight to Traj. Length Ratio	5.6	29	Avg. Velocity	18.2	10	SC Moving Rotation	3.3	47
Avg. Distance	8.7	24	Width	13.4	22	FC Straight Length	18.3	7
Avg. Size	15.5	9	Height	15	19	SC Straight Length	16.8	11
Avg. Pressure	14	13	Area Size	15.8	18	Straight Length Ratio	16.8	12
Distance STD Deviation	4.9	31	Width-to-Height Ratio	10.7	32	FC Traj. Radius	5.3	45
Size STD Deviation	9.1	23	Size STD Deviation	12.6	25	SC Traj. Radius	1.7	49
Pressure STD Deviation	9.7	20	Pressure STD Deviation	10.3	34	Avg. Size of FC	15.5	19
						Avg. Size of SC	17.1	9
FC Pressure STD Deviation	7.7	42	FC Size STD Deviation	11	31	Avg. Pressure of FC	12.2	26
SC Pressure STD Deviation	9.8	36	SC Size STD Deviation	10.2	33	Avg. Pressure of SC	11.4	28

3.2 Classification

The major purpose of the classification process in Fig. 1 is to authenticate users using a classifier. We discuss our *authentication model* and classifier in this section. Moreover, since there is no systematic study of touch biometric properties so far, we further introduce our *discrimination model* for studying its biometric properties. The key difference of a discrimination model from an authentication model is that, in a discrimination model, we can have the data of each class for training. Fig. 4 compares these two models and visualizes their difference.

3.2.1 Discrimination Model

We define this model as a typical multi-class classification model: Given N classes, each having some samples, how to identify which one of these classes a new observation belongs to. In the training phase, a number of labeled touch data from N users are processed via the feature extraction process discussed in Section 3.1. We can obtain corresponding N classes of feature vectors. The vectors are then fed into a classifier for training purpose. While in the discrimination phase, a new touch data observation is also processed via feature extraction process first. The classifier then decides which class the obtained feature vector belongs to and then identify the user accordingly.

Obviously, when N grows, the identification process naturally becomes more difficult, and the accuracy would decrease. A form of good biometrics should exhibit good performance even when N is large. Hence, the discrimination model can reflect the distinctiveness of biometric properties by involving different numbers of users.

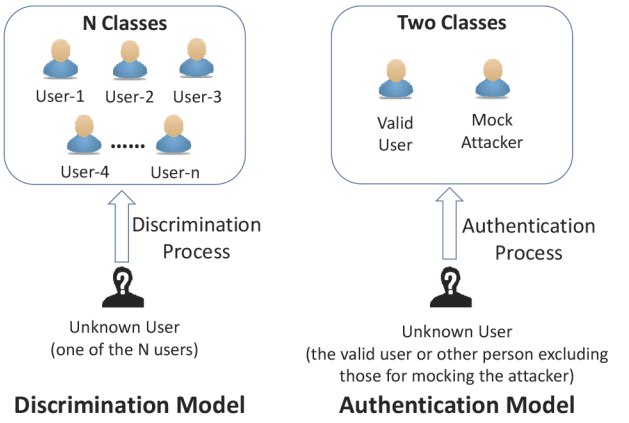


Figure 4: Comparison between discrimination model and authentication model

3.2.2 *Authentication Model*

In practice, we cannot know the models of the attackers beforehand. However, we can obtain the touch data of the valid user herself, and those of some other users[3]. We use these additional users to build a mock attacker model as an approximation to the real, unknown attacker.

We define the authentication problem as a binary classification problem. Given two classes of samples, one including touch data of the valid user, and the other including those of the mock attackers, how to identify which class a *new* observation belongs to. In the training phase, given the touch data of both classes, we can obtain two corresponding classes of feature vectors via the feature extraction process discussed in Section 3.1. We can then turn to a classification algorithm: Input the two classes of feature vectors to train a classifier. After the classifier is trained, it can be used to determine whether a current user operation is from a valid user or not, by checking which class (*i.e.*, the valid user class or the attacker class) it belongs to.

3.2.3 *Classifier*

There are many classification algorithms we can choose. We adopt a state-of-the-art statistics-based classification method, *i.e.,* the Support Vector Machine (SVM) [6]. It can infer how two classes of vectors are different from each other by finding a hyperplane (*i.e.*, a boundary) that best separates the classes. With such a boundary, any unlabeled sample can then be classified according to which side of the boundary it locates.

We adopt SVM since it has long been proven successful in many classification applications. Moreover, it can seamlessly apply the kernel method, *e.g.*, via Radial Basis Function (RBF) kernel [6], and thus find a nonlinear boundary that best separates the two classes. This non-linear property is critical to our problem setting, since the discriminations between the touch data from the valid user and those from the attackers are nonlinear in nature.

Finally, note that SVM is not the only option of classifier for our user authentication approach. Other methods, for example, logistic regression and Naive Bayes classifier, can also be incorporated into our approach conveniently. A further comparison study is left to our future work.

[3]These data are collectable in reality since it is not hard to collect the touch data of some other users who use the same smartphone model.

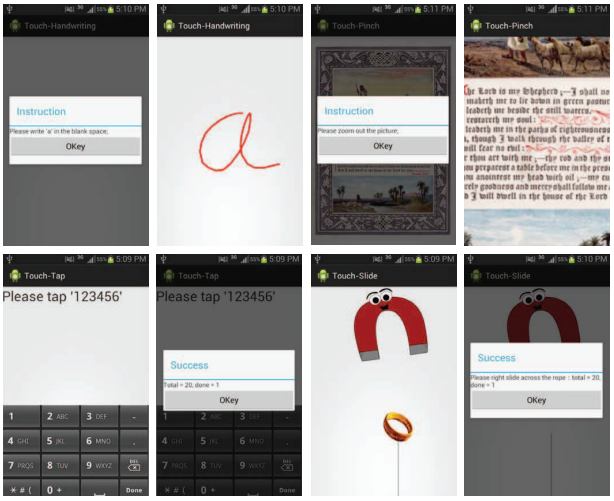


Figure 5: User interface of our data acquisition tool. The first row demonstrates our handwriting and pinch experimental UIs, while the other demonstrates these of keystroke and slide.

4. EXPERIMENTAL STUDY

In the previous section, we have described our framework for continuous authentication based on touch operations. This section evaluates its performance via real-world experiments. First, we conduct a real-world experiment to collect touch data. Secondly, we evaluate the proposed features using these data. Thirdly, we study the distinctiveness and permanence properties of touch operation, and justify it qualifies to be a form of good biometrics. Finally, we evaluate the authentication performance of our proposed framework.

4.1 Data acquisition

We recruited 32 participants for our data acquisition experiment using an online advertisement. The only requirement was that the participants had to be users of smartphone with touchscreen. This was to guarantee that they were familiar to the touch operations required in the experiment. Each participant received a $6 gift for his/her participation.

In order to collect touch data, we programmed a data acquisition tool with Java, which runs on Android smartphone as a stand-alone application. This tool collects the four types of touch operations of interest, and saves their touch data sequences for further analysis. Fig. 5 shows the user interface of this tool. It was installed on a Samsung Galaxy SII smartphone with Android OS 4.1.2.

Before the experiment, the participants were informed that that their touch data would be collected for behavior analysis, and they were required to operate as they usually did. After they got familiar with the tool, we required them to start performing operations as the tool instructed. Each experiment took roughly 15 minutes. In this way, we collected 200 touch data sequences from each participant.

We further chose 3 volunteers among these 32 participants for a long-term study. We asked them to do the experiment

with the same settings repeatedly for 20 more times. The interval of each two consecutive experiments for each volunteer was one day by default except weekends. To be convenient, we only required them to perform tasks for about 5 minutes (*i.e.*, we thus collected 50 touch data sequences) in each experiment. The whole data acquisition experiment lasted for almost one month. We collected roughly 1200 touch data sequences from each volunteer in total[4].

4.2 Feature Evaluation

In Section 3.1, we have suggested a set of features for each type of touch operations. We now evaluate the effectiveness of each feature in classification accuracy. The idea is to discriminate users *solely* based on one feature at a time. We adopt the discrimination model in the feature evaluation process. To elaborate, in the training phase, we use only *one* feature to model the user at a time. The classifier then classifies a new sample based on this model. The classification accuracy can be obtained accordingly as an indication of the feature's effectiveness.

In our experimental settings, we use the data set of 32 participants. To evaluate each feature, the classifier performs a 10-fold cross validation based on the data of that particular feature. A 10-fold cross validation approach randomly partitions the data into 10 equal-size subsets. Each time nine subsets are used for training, and the remaining subset is retained for testing. The accuracy values are then averaged. Our evaluation results are provided in Table 2 along with the feature name, and the ranking according to the accuracy.

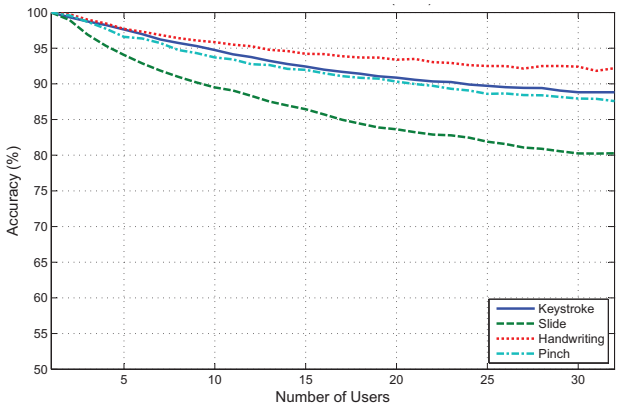
According to [15], a feature X is relevant in the process of discriminating class $Y=y$ from others if the conditional probability $P(Y=y|X=x)$ is different from the unconditional probability $P(Y=y)$ for some values $X=x$ for which $P(X=x)>0$. In our study, since the task is to discriminate one user among the 32 users, a naive guess can achieve a 1/32 accuracy (*i.e.*, the unconditional probability is 3.125%). Therefore, the features with accuracy lower than 3.125% are useless in discriminating users, and we thus remove these features.

In the rest of our study, we consider only the features with accuracy higher than 3.125% in Table 2. Noticing that some features on directional information are not discriminating. We believe such an evaluation study can enlighten future feature extraction method for touch-based continuous authentication.

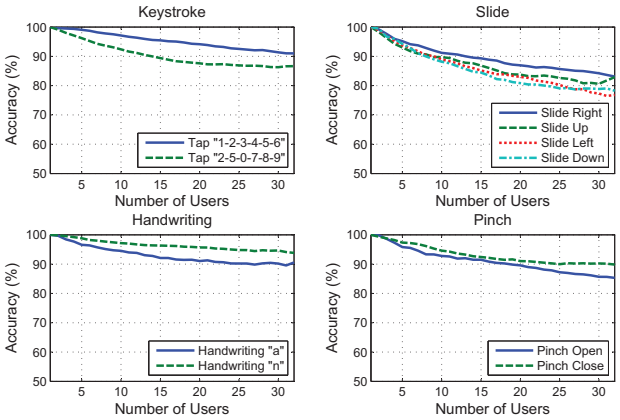
4.3 Evaluation of Distinctiveness

In this section, we evaluate the distinctiveness of touch biometrics, *i.e.*, how well touch operations can be used to discriminate users. We adopt the discrimination model in this step. Our experiment is based on the data set of feature vectors from 32 users. We randomly pick N users and their vectors from the data set. Focusing on each type of touch operation at a time, we benchmark the classification accuracy with N users using a 10-fold cross validation approach. We change N from 2 to 32, and thus get the accuracy with different user sizes. Fig. 6(a) shows our experiment results. We can see that all types of touch operations are distinctive among users with a classification accuracy better than 80% even when we try to discriminate a user from 31 others.

[4]The data set are available at the project homepage: http://www.cudroid.com/urmajesty.



(a) Overall distinctiveness performance



(b) Distinctiveness performance of touch operation subtypes

Figure 6: Distinctiveness performance of touch operations based on the data set of 32 users

We have noticed that there are still minor differences among the operations of each type. Specifically, a pinch may be pinch open or pinch close; A slide can have four possible directions; Handwriting can involve different letters; Keystroke operations can input different words. We study whether such *subtypes* have a considerable impact on the distinctiveness performance. Fig. 6(b) shows the experiment results, from which we can tell that the differences between subtypes are slight. Therefore, in the subsequent experiments, we will not consider these subtypes.

4.4 Evaluation of Permanence

We now study the permanence performance of touch biometrics, *i.e.*, if we model a user with her touch biometrics, whether the model is stable over a period of time for the same user. In this regard, our experiment is based on a 21-day data set from the 3 volunteers. As mentioned before, we collected their touch data from a 21-day long experiment. We use the discrimination model for evaluation. To elaborate, we model the user using their data collected in the first day. We then discriminate the data of each remaining day based on this model. If touch biometrics bears good permanence property, the model should be good enough in discriminating the data of the remaining days. Fig. 7 shows the results.

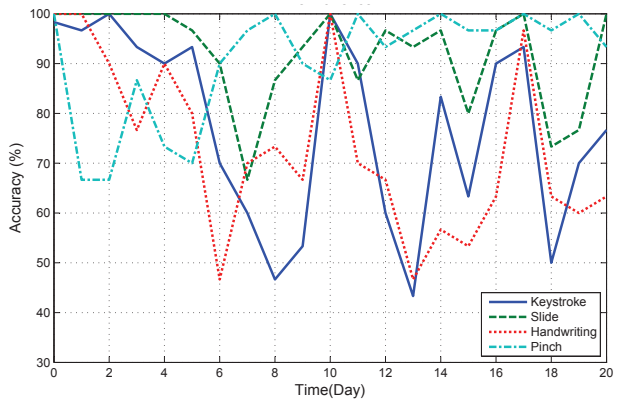


Figure 7: Permanence performance based on the data set of the 3 volunteers

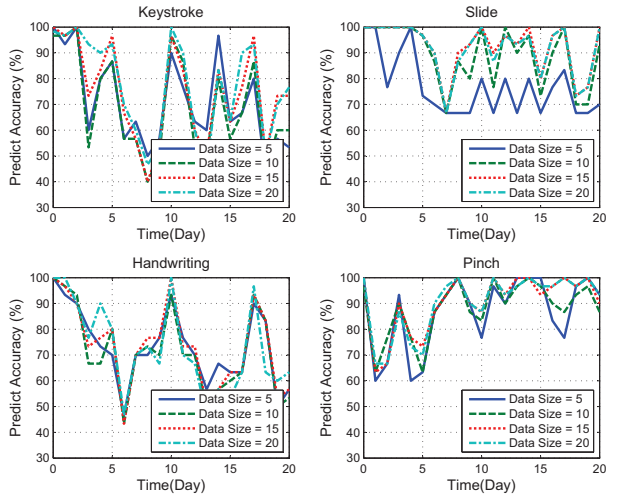


Figure 8: Permanence performance with different training data sizes

We can observe that the performance is not stable for all touch operations, even though pinch and slide are relatively better than keystroke and handwriting. It is probably because that our data used for training is too flaky to get a stable enough result. To further clarify this issue, we conduct another experiment using different sizes of training samples. The results in Fig. 8 show that the performance improves only a little as the data size grows. Therefore, we can infer that data size is not the key factor to the poor performance. As a result, we conclude that touch biometrics is not quite stable over time.

A common way to deal with the permanence issue in biometric systems is to consider an adaptive approach: The model will be adjusted according to new samples. We investigate whether such an adaptive approach is helpful for touch biometrics. For this reason, we improve the previous experiment in permanence evaluation using the same data set. When discriminating the data of the nth day, we model the users using all the touch data previous to the nth days, instead of the first day only. Fig. 9 shows the evaluation results. We can see that the results tend to be much more stable, especially after the 8th day. This shows that an adaptive approach can help tackle the permanence problem.

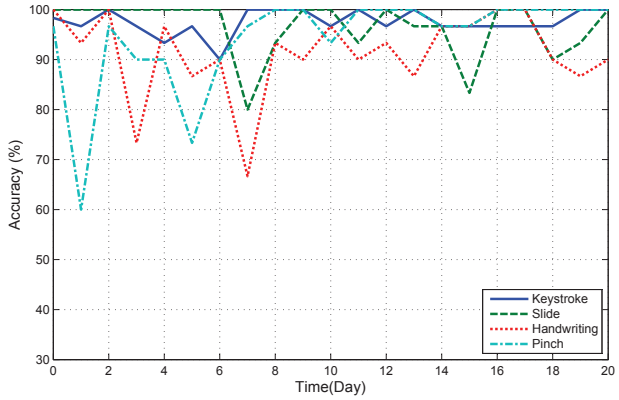


Figure 9: Permanence performance of adaptive approach based on the data set of the 3 volunteers

Table 3: Average error rate with different numbers of additional users to model the mock attacker

additional user #	Keystroke	Slide	Handwriting	Pinch
5	11.76%	11.24%	11.48%	7.38%
10	10.3%	10%	10.08%	4.96%
15	9.36%	4.85%	9.27%	3.87%
20	7.71%	1.53%	11.39%	3.75%
28	6.42%	0.75%	8.67%	3.33%
30	5.3%	1.3%	8.67%	3.33%

4.5 Evaluation of Touch-based Authentication

In this section, we study the performance of touch-based authentication. The major difference of this study is that we consider the practical case, where the attacker model is not known beforehand. In other words, the classifier cannot be trained with the touch data from the real attacker. We adopt the authentication model in this study. As discussed in Section 3.2.2, we assume that we can have the touch data of the valid user herself, and those of some other users to mock attackers.

Our experimental setting is discussed as follows. We consider each of the 3 volunteers at a time, and use her data of the previous 20 days to model the valid user. We then randomly select M additional users from the rest 31 users to model the mock attacker. The remaining data of the valid user and those of the rest users (those are not involved in the training process) are used for prediction. We study the performance in terms of average error rate (*i.e.*, (FAR+FRR)/2). Table 3 shows our experiment results when M varies. Each error rate within this table is an average of those of the three volunteers'.

From Table 3, we can observe that the performance improves as the additional users number increases. However, an overfitting for slide occurs when the number of additional users exceeds 28. But for the other 3 touch operations, the performance might further improve when involving more additional users.

In general, including more additional users can help reduce FAR, since it explores more diverse user characteristics. In other words, involving more additional users shrink the class boundary of the valid user and thus improve FAR.

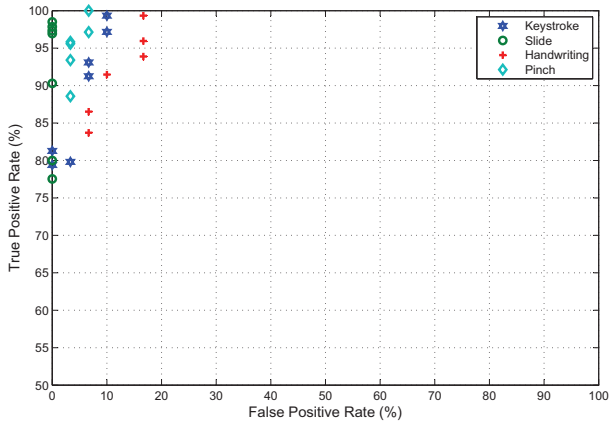


Figure 10: ROC plot when using different number of additional users to model the mock attacker.

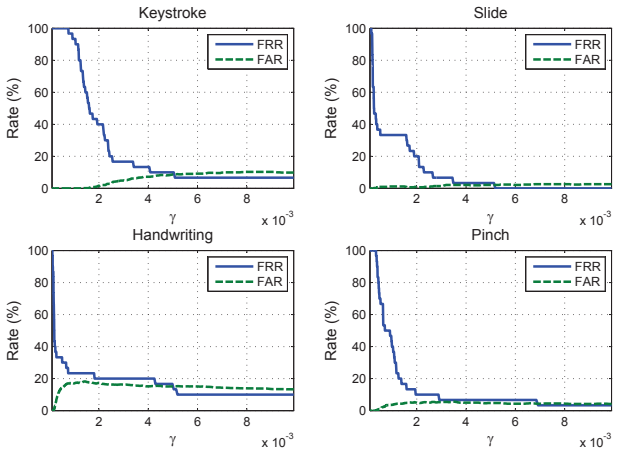


Figure 11: FAR/FRR plots where additional user number equals to 20 for keystroke, 25 for slide, 8 for handwriting and 20 for pinch

However, when the number of additional users are too high (*e.g.*, the 30 case), it may also deteriorate the authentication accuracy. This is not surprising: As the number increases, the attacker samples are getting more diverse, and the SVM will suffer overfitting to the attacker class. As a result, it tends to misclassify more operations of the valid users, causing a high FRR.

In practice, FAR and FRR are correlated with each other. To avoid bias, ROC is commonly used to evaluate biometric systems, which reflects the characterization of the trade-off between the true positive rate and the false positive rate. Fig. 10 visualizes such a trade-off for the average error rate achieved in Table 3.

Since our approach heavily relies on the SVM classifier, we tune the SVM parameters to get the EER. We adopt a commonly-used RBF kernel in the SVM classifier, defined as $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$ [6]. We tune the value of γ and obtain corresponding FAR and FRR, which are plotted in Fig. 11. We observe that our biometric system can generally achieve EER values lower than 10% for all operation types. The slide operation performs the best by achieving an EER lower than 1%.

Table 4: Average error rate using consecutive sequences. To better visualize the improvement, some previous experiment results in Table 3 are also shown here for comparison purpose.

user # in training	Numbers of Operation		
	1	3	5
Keystroke			
10	10.3%	9.82%	9.71%
20	7.7083%	7.74%	3.32%
28	6.4167%	5.02%	0.88%
Slide			
10	10%	9.55%	9.33%
20	1.5278%	0.98%	0.64%
28	0.75%	0%	0%
Handwriting			
10	10.0758%	5.94%	5.62%
20	11.3889%	10.92%	15.8%
28	8.6667%	8.3%	13.89%
Pinch			
10	4.9621%	2.63%	2.1%
20	3.75%	1.47%	0.92%
28	3.333%	0%	0%

In practical scenarios, we can use a combination of consecutive operations jointly for making an authentication decision [14]. A convenient approach is to authenticate the user with each of the operations first. The system then decides whether a user is an attacker based on the majority of the results. To verify the applicability of this idea to our model, we conduct a comparison experiment with the same data set. This time, we try to authenticate users with 3 and 5 consecutive operations. Table 4 shows the experiment results, which confirm such an approach is helpful in improving authentication performance. According to Table 4, the performance improves a lot in most cases. For slide and pinch, the average error rate even approaches 0. However, the performance for handwriting does not improve much. We think the reason is that the average error rate for each handwriting operation is relatively high. From the permanence experiment, we could infer that consecutive handwriting operations are more likely to be similar. Therefore, errors would also tend to happen consecutively in a short interval, rather than distribute evenly over a period of time. When such case occurs, the performance will degrade due to the high error rate. Which will affect the performance when the error rate is too high. If the rate could be lower down (*e.g.*, by involving more additional users), the result would also improve. Details of such an evaluation are left to future work.

To conclude, when we model the mock attacker properly, the authentication performance can be very promising. Also, using consecutive sequences to authenticate a user is a helpful way to improving the error rate.

4.6 Lessons Learned

Our experiments have evaluated the distinctiveness and permanence properties of touch operations. The results show that touch operation can be a form of good biometrics. However, regarding the distinctiveness property, we find that there is still room for the accuracy to approach 100% when we discriminate the users. As a result, our touch-

based continuous authentication approach cannot achieve an error rate very close to zero using one operation. This indicates a need for further research to make touch-based continuous authentication a practical solution. We believe that it is a promising solution to consider a set of touch operations jointly for making an authentication decision rather than using one at a time. We have shown that when considering 3 or 5 consecutive operations jointly, the biometric system achieves average error rates approaching 0% for slide or pinch, which can satisfy practical concerns. However, how to use these operation combinations effectively and efficiently should be studied in the future.

Regarding the permanence property, we find that touch biometrics are not strictly stable over time, especially for keystroke and handwriting. We have shown that a convenient adaptive approach can greatly improve the accuracy. Therefore, the permanence problem can be mitigated. However, a more sophisticated approach is still at large.

Finally, touch-based authentication inevitably requires a large number of touch operation samples for training purpose. We have shown that potential attackers can be modeled with data from a set of additional users. Such data can be preloaded into smartphone in practice. However, what is the adequate number of additional users should be further studied in the future. Moreover, we still need hundreds of training samples from the target valid user. How to design a user-friendly way to obtain so many data samples is still an open question for implementing touch-based authentication.

5. RELATED WORK

Continuous authentication on traditional PC has been extensively studied for years. Research on how to continuously authenticate PC users can be found in [2, 7, 18, 19, 20, 28, 30, 36]. Keystrokes, mouse dynamics, and face recognition are the main approaches. However, the usability of these technologies is still a question due to the low recognition accuracy and inconvenience.

Equipped with more sensors in smartphones (*e.g.*, gyroscopes), continuous authentication on smartphone started a new research area. Several projects have studied how to passively authenticate users based on a variety of sensory data. For example, SenSec [38] constantly collects sensory data from accelerometers, gyroscopes and magnetometers, and constructs the gesture model of how a user uses the device. The user studies has showed that SenSec achieved an accuracy of 75% in identifying the users and 71.3% in detecting the non-owners. Senguard [29] also investigates on a framework to continuously identify users based on a variety of sensory data. Touchscreen is one sensor of concern. However, the paper only visually shows that different users have different touch traces, without mentioning how to authenticate users based on these traces.

Using touch operations to authenticate users is a relatively new topic that has yet to capture extensive research attentions. Several recent work has studied how to improve the touch unlocking mechanism by considering touch biometrics. Such work includes [3, 9, 25, 26, 33]. De Luca *et al.* in [9] propose to track touch data of slide operations to unlock the screen. Touch data including time, position, size and pressure are used directly to authenticate users. Their work has achieved an overall accuracy of 77% using DTW (*i.e.*, Dynamic Time Warping) at best. Angulo *et al.* research on improving the lock patterns and introduce the notion of lock pattern dynamics [3]. Their work has achieved an EER of 10.39% using Random Forest machine learning classifier. Sae-Bae *et al.* focus on the specific five-finger touch gestures available on the Apple devices [25]. They model a user based on the movement characteristics of the five fingers and the palm center. An accuracy of 90% has been achieved over an Apple iPad. Shahzad *et al.* discuss a slide-based user authentication scheme, where a series of customized slides are used jointly to authenticate users [26]. It has been reported that a combination of three slides can achieve an average EER of 0.5%. Sun *et al.* propose TouchIn that allows user to draw on arbitrary regions with one or multiple fingers to unlock his mobile device. The user is authenticated based on the geometric properties of his drawn curves as well as his behavioral and physiological characteristics [33].

Other than improving screen locker security, several investigations focus on exploring the applicability of traditional keystroke-based authentication on smartphone with new features. KenSens [10] passively authenticates users via the specific location touched on each key, the drift from finger down to finger up, the force of touch, the area of press. The work in [23] also discusses the feasibility of employing keystroke dynamics to perform user verification on mobile phones and introduces a new statistical classifier. However, such work has not achieved great improvement in authentication accuracy. Zheng *et al.* propose to rely on more sensors (*e.g., accelerometers*) other than purely touchscreen [37]. They propose acceleration features which can reflect the magnitude of acceleration when the key is pressed and released. Their approach finally has achieved an average EER down to 3.65%.

Besides exploring touching biometrics on improving the screen lock or keystrokes, Frank *et al.* introduce the notion of continuous authentication via touch operations [14]. They focus on stroke operations. An EER of 13% for one single stroke, and 2% to 3 % for 11 consequent strokes have been achieved. Instead of only considering slide operation, Li *et al.* study both tap and slide, and achieved an accuracy of approximately 90% [21]. Feng *et al.* also study the continuous mobile authentication issues via touchscreen gestures [12]. They implement FAST (*i.e.*, Finger-gestures Authentication System using Touchscreen), where an extra glove equipped with sensors is used. FAST has achieved an FAR of 4.66% and an FRR of 0.13% using 7 touch sequences.

Our work also aims at exploring the applicability of continuous authentication relying only on touch operations. Unlike the existing work that using only one type of specific touch operation, our work comprehensively investigates a set of general, commonly-used types of touch operations on smartphone. Our authentication performance is better than that reported in [14] and [21] (the other existing work focuses on different problem settings, and is not comparable). More importantly, all existing work is based on the hypothesis that touch data qualifies good biometrics. Our work is the first to systematically evaluate the distinctiveness and permanence properties of touch biometrics. Such a study is the basis for touch-based authentication.

6. CONCLUSION

This work has suggested a touch-based authentication framework to continuously authenticate user. The authentication proceeds in a passive way while the user performs her normal touch operations. We proposed a set of meth-

ods targeting the problem of how to model multiple types of touch data produced by users. We further justified two critical properties of such data: distinctiveness and permanence. We presented our work together with a real-world experimental study. It is the first attempt to comprehensively evaluate the biometric properties of touch operations.

Although we have shown that touch operations bear good biometric properties, there is still a long way to implement a practical, touch-based continuous authentication system. First, the error rate when authenticating a user with one touch operation still cannot approach zero. We have hence suggested considering a set of touch operations jointly. Although we have shown some preliminary results with such a consideration, future research efforts (*e.g.*, consider the combination of different touch operations) are still required to examine it comprehensively. Secondly, our experiments have shown that the user features of touch operations are not stable over a period of time. Although we have suggested an adaptive approach that can mitigate such a problem, extensive future work is still needed to find an optimized adaptation method. Finally, there are quite a lot of other implementation issues of our touch-based continuous authentication framework. Examples include how to engineer a seamless touch operation tracing mechanism that runs silently as a smartphone background service and how to design a user-friendly mechanism to obtain data samples for training purpose.

Acknowledgements

The authors are grateful to our shepherd, Dr. Alex De Luca, who helps a lot in improving this manuscript. We also thank Xiaolei Zhang and the anonymous reviewers for their constructive comments. This work was supported by the Key Project of National Natural Science Foundation of China (Project No. 61332010), the National Basic Research Program of China (973 Project No. 2014CB347701), the National Natural Science Foundation of China (Project No. 61100077), the Shenzhen Basic Research Program (Project No. JCYJ20120619152636275), and the Research Grants Council of Hong Kong (Project No. CUHK 415113). Yangfan Zhou is the corresponding author.

7. REFERENCES

[1] Receiver operating characteristic. *http://en.wikipedia.org/wiki/Receiver_operating_characteristic*.

[2] A. Altinok and M. Turk. Temporal integration for continuous multimodal biometrics. In *Proc. of the Workshop on Multimodal User Authentication*, 2003.

[3] J. Angulo and E. Wästlund. Exploring touch-screen biometrics for user identification on smart phones. In *Privacy and Identity Management for Life*, pages 130–143. Springer, 2012.

[4] L. C. Araújo, L. H. Sucupira Jr, M. G. Lizarraga, L. L. Ling, and J. B. T. Yabu-Uti. User authentication through typing biometrics features. *IEEE Trans. on Signal Processing*, 53(2), Feb. 2005.

[5] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith. Smudge attacks on smartphone touch screens. In *Proc. of the 4th USENIX Conf. on Offensive Technologies*, pages 1–7, 2010.

[6] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[7] I. Brosso, A. La Neve, G. Bressan, and W. Ruggiero. A continuous authentication system based on user behavior analysis. In *Proc. of the 10th Int. Conf. on Availability, Reliability, and Security*, 2010.

[8] E. Chin, A. P. Felt, V. Sekar, and D. Wagner. Measuring user confidence in smartphone security and privacy. In *Proc. of the 8th Symposium on Usable Privacy and Security*, 2012.

[9] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. Touch me once and i know it's you! implicit authentication based on touch screen patterns. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2012.

[10] B. Draffin, J. Zhu, and J. Zhang. Keysens: passive user authentication through micro-behavior modeling of soft keyboard interaction. In *Proc. of the 5th Int. Conf. on Mobile Computing, Applications and Services*, 2013.

[11] Egham. Gartner says smartphone sales grew 46.5 percent in second quarter of 2013 and exceeded feature phone sales for first time. *http://www.gartner.com/newsroom/id/2573415*, 2013.

[12] T. Feng, Z. Liu, K.-A. Kwon, W. Shi, B. Carbunar, Y. Jiang, and N. Nguyen. Continuous mobile authentication using touchscreen gestures. In *Proc. of the IEEE 6th Int. Conf. on Biometrics: Theory, Applications and Systems*, 2013.

[13] I. Fischer, C. Kuo, L. Huang, and M. Frank. Smartphones: not smart enough? In *Proc. of the 2nd ACM workshop on Security and privacy in smartphones and mobile devices*, Oct. 2012.

[14] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. Touchalytics: on the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE Trans. on Information Forensics and Security*, 8(1), Jan. 2013.

[15] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*. Springer-Verlag, 2006.

[16] A. K. Jain, A. Ross, and S. Pankanti. Biometrics: a tool for information security. *IEEE Trans. on Information Forensics and Security*, 1(2), June 2006.

[17] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(1), Jan. 2004.

[18] R. Janakiraman, S. Kumar, S. Zhang, and T. Sim. Using continuous face verification to improve desktop security. In *Proc. of the 7th IEEE Workshops on Application of Computer Vision*, 2005.

[19] A. J. Klosterman and G. R. Ganger. Secure continuous biometric-enhanced authentication(cmu-cs-00-134). *CMU Technical Report*, 2000.

[20] G. Kwang, R. H. C. Yap, T. Sim, and R. Ramnath. An usability study of continuous biometrics authentication. *Advances in Biometrics*, (828-837), 2009.

[21] L. Li, X. Zhao, and G. Xue. Unobservable reauthentication for smart phones. In *Proc. of the 20th Network and Distributed System Security Symposium*, volume 13, 2013.

[22] D. T. Lin. Computer-access authentication with

neural network based keystroke identity verification. In *Proc. of the Int. Conf. on Neural Networks*, 1997.

[23] E. Maiorana, P. Campisi, N. González-Carballo, and A. Neri. Keystroke dynamics authentication for mobile phones. In *Proc. of the 2011 ACM Symposium on Applied Computing*, 2011.

[24] C. E. Metz. Basic principles of roc analysis. In *Seminars in Nuclear Medicine*, volume 8, pages 283–298. Elsevier, 1978.

[25] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon. Biometric-rich gestures: a novel approach to authentication on multi-touch devices. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2012.

[26] M. Shahzad, A. X. Liu, and A. Samuel. Secure unlocking of mobile touch screen devices by simple gestures: you can see it but you can not do it. In *Proc. of the 19th Annual Int. Conf. on Mobile Computing and Networking*, pages 39–50, 2013.

[27] C. Shen, Z. Cai, and X. Guan. Continuous authentication for mouse dynamics: a pattern-growth approach. In *Proc. of the 42nd Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks*, June 2012.

[28] S. J. Shepherd. Continuous authentication by analysis of keyboard typing characteristics. In *Proc. of the European Convention on Security and Detection*, 1995.

[29] W. Shi, J. Yang, Y. Jiang, F. Yang, and Y. Xiong. Senguard: passive user identification on smartphones using multiple sensors. In *Proc. of the 7th Int. Conf. on Wireless and Mobile Computing, Networking and Communications*, 2011.

[30] T. Sim, S. Zhang, R. Janakiraman, and S. Kumar. Continuous verification using multimodal biometrics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4), Apr. 2007.

[31] A. Smith. Smartphone ownership (2013 update). *http://pewinternet.org/Reports/2013/Smartphone-Ownership-2013.aspx*.

[32] S. N. Srihari, S. H. Cha, H. Arora, and S. Lee. Individuality of handwriting. *Journal of Forensic Sciences*, 47(4), July 2002.

[33] J. Sun, R. Zhang, J. Zhang, and Y. Zhang. Touchin: Sightless two-factor authentication on multi-touch mobile devices. *http://arxiv.org/abs/1402.1216*, 2014.

[34] S. Thomas. Touchscreen handsets dominanting uk mobile market! *http://www.3g.co.uk/PR/Nov2012/touchscreen-handsets-dominanting-uk-mobile-market.html*, 2012.

[35] D. Van Bruggen, S. Liu, M. Kajzer, A. Striegel, C. R. Crowell, and J. D'Arcy. Modifying smartphone user locking behavior. In *Proc. of the 9th Symposium on Usable Privacy and Security*, 2013.

[36] R. H. C. Yap, T. Sim, G. X. Y. Kwang, and R. Ramnath. Physical access protection using continuous authentication. In *Proc. of the IEEE Conf. on Technologies for Homeland Security*, 2008.

[37] N. Zheng, K. Bai, H. Huang, and H. Wang. You are how you touch: user verification on smartphones via tapping behaviors(wm-cs-2012-06). *Tech. Repo. of the College of William and Mary*, 2012.

[38] J. Zhu, P. Wu, X. Wang, and J. Zhang. Sensec: mobile security through passive sensing. In *Proc of the 13th Int. Conf. on Computing, Networking and Communications*, 2013.

Modeling Users' Mobile App Privacy Preferences: Restoring Usability in a Sea of Permission Settings

Jialiu Lin Bin Liu Norman Sadeh Jason I. Hong
School of Computer Science, Carnegie Mellon University
{jialiul, bliu1, sadeh, jasonh}@cs.cmu.edu

ABSTRACT

In this paper, we investigate the feasibility of identifying a small set of privacy profiles as a way of helping users manage their mobile app privacy preferences. Our analysis does not limit itself to looking at permissions people feel comfortable granting to an app. Instead it relies on static code analysis to determine the purpose for which an app requests each of its permissions, distinguishing for instance between apps relying on particular permissions to deliver their core functionality and apps requesting these permissions to share information with advertising networks or social networks. Using privacy preferences that reflect people's comfort with the purpose for which different apps request their permissions, we use clustering techniques to identify privacy profiles. A major contribution of this work is to show that, while people's mobile app privacy preferences are diverse, it is possible to identify a small number of privacy profiles that collectively do a good job at capturing these diverse preferences.

1. INTRODUCTION

As of December 2013, the Google Play Store offered more than 1,130,000 apps; the Apple App store offered more than 1,000,000 apps. Each store has reported more than 50 billion downloads since its launch [1, 2]. The growth in the number mobile apps has in part been fueled by the increasing number APIs made available to developers, including a number of APIs to access sensitive information such as a user's current location or call logs. While these new APIs open the door to exciting new applications, they also give rise to new types of security and privacy risks. Malware is an obvious problem [3, 4]; another danger is that users are often unaware of how much information these apps access and for what purpose.

Early studies in this area have shown that privacy interfaces, whether for iOS or for Android, did not provide users with adequate information or control [5-7]. This was quickly followed by research exploring solutions that offered users finer grain control over the use of these APIs [8-10]. Perhaps because of this research, iOS and Android have now started to offer their users somewhat finer control over mobile app permissions, enabling them for instance to toggle permissions on and off on an app-by-app basis (e.g. iOS5 and above, and also App Ops in Android 4.3). However, with users having an average of over 40 apps on their smartphone [11] and each app requiring an average of a little over 3 permissions [12], systematically configuring all these settings places an unrealistically high burden on users.

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9-11, 2014, Menlo Park, CA.

This paper investigates the feasibility of organizing end-users into a small set of clusters and of identifying default privacy profiles for each such cluster as a way of both simplifying and enhancing mobile app privacy. We use data obtained through static code analysis and crowdsourcing, and analyze it using machine learning techniques to highlight the limitations of today's interfaces as well as opportunities for significantly improving them. Specifically, our results were obtained by collecting 21,657 preference ratings from 725 users on 837 free Android apps. These preference ratings were collected on over 1200 app-permission-purpose triples. Each such preference rating captures a user's willingness to grant a given permission to a given app for a particular purpose. Identification of the purpose(s) associated with a given app's permission was inferred using static code analysis, while distinguishing between different types of 3rd-party libraries responsible for requesting access to a given permission. For example, if location data is used by an app only because of an ad library bundled with the app, we can infer that location is used for advertising purposes.

Our analysis indicates that a user's willingness to grant a given permission to a given mobile app is strongly influenced by the purpose associated with such a permission. For instance a user's willingness to grant access to his or her location will vary based on whether the request is required to support the app's core functionality or whether it is to share this information with an advertising network or an analytics company. Our analysis further shows that, as in many other privacy domains, people's mobile app privacy preferences are diverse and cannot adequately be captured by one-size-fits-all default settings. Yet, we show that it is possible to cluster users into a small number of privacy profiles, which collectively go a long way in capturing the diverse preferences of the entire population. This in turn offers the prospect of empowering users to better control their mobile app permissions without requiring them to tediously review each and every app-purpose-permission for the apps on their smartphones. Beyond just mobile apps, these results open the door to privacy interfaces that could help reconcile tensions between privacy and user burden in a variety of domains, in which explosion in functionality and usage scenarios are stretching demands on users (e.g. browser privacy settings, Facebook settings, and more).

The contribution of this research is threefold. First, we provide an in-depth analysis of mobile app permissions that is not limited to the types of sensitive resources an app requests (e.g. location, contact lists, account information) but also includes the "*purpose*" associated with these requests – with purpose identified through static analysis of third party libraries and their API calls. Second, we describe the results of a larger-scale version of the crowdsourcing methodology originally introduced by Lin et. al. [13]), collecting over 21,000 privacy preferences associated with different permissions and purposes. This allows us to quantitatively link users' mobile app preferences to different

types of app behaviors that involve sensitive resource usage. Third, we present a clustering analysis of the privacy preferences of 725 smartphone users, and show that, while these preferences are diverse, a relatively small number of privacy profiles can go a long way in simplifying the number of decisions users have to make. This last contribution offers the promise of alleviating user burden and ultimately increasing their control over their information.

2. RELATED WORK

A great deal of past work analyzing smartphone apps has focused on developing useful techniques and tools to detect and manage leakage of sensitive personal information [8-10, 14-26] or studying how users react to these usages [6, 13, 27, 28]. In this section, we summarize the relevant mobile privacy literature, which we organize around three themes.

2.1 Finer Grain Privacy Controls

In Android, apps can only access sensitive resources if they declare permission requests in manifest files[1] and obtain authorization from users to access these permissions at download time. Several studies have examined usability issues related to the permission interface displayed to users as they download Android apps [5-7]. The studies have shown that Android permission screens generally lack adequate information, with most users struggling to understand key terms and the implications associated with the permissions they are requested to grant.

Android 4.3 saw the introduction of a hidden permission manager referred to as a "App Ops" that allows users to review and manipulate settings associated with the permissions of the apps they have downloaded on their smartphones [29, 30], This feature was later removed in Android 4.4 presumably due to usability problems – namely the unrealistically large number of permission decisions already mentioned in Section 1. Similar fine grain control over permissions has also been offered by third party privacy manager apps, such as LBE privacy guard [31], though it is only available on rooted Android devices. Similar settings are also available in iOS (iOS 5 and above), where users have the ability to turn on and off access to sensitive data or functionality (such as location, contacts, calendars, photos, etc) on an app-by-app basis. ProtectMyPrivacy [32] offers similar settings to jailbroken iPhone users and also provides recommendations based on majority voting (effectively looking for popular one-size-fits-all settings, when such settings can be identified).

A number of research prototypes have also offered used fine grain controls over the permissions [8, 10, 32-35]. MockDroid [8] and TISSA [10] also allow users to ibject fake information in response to API calls made by apps. AppFence [9], a follow-up to TaintDroid [17], also allows users to specify resources, which should only be used locally. Apex proposed by Nauman et al. [34] provides fine-grained control over resource usage based on context and runtime constraints.

These proposed privacy extensions aim to provide users with finer control over the data accessed by their apps. However, these extensions also assume that users can correctly configure all the resulting settings. We argue that asking users to specify such a large number of privacy preferences is unrealistic. In addition, we show that controlling permissions on an app-by-app basis without taking into account the purpose of these permissions does not enable one to capture important differences in people's mobile app privacy preferences. The present paper complements prior work in this area by identifying a small number of manageable privacy profiles that takes into account purpose and offers the promise of empowering users to manage their mobile app privacy without imposing an undue burden on them.

2.2 Modeling People's Mobile App Privacy Preferences

A second line of research has focused on studying users' mobile app privacy concerns and preferences. For example, Felt et al. [28], Chin et al. [27], and Egelman et al [36] conducted surveys and interviews to understand mobile users' mobile privacy concerns as well as their over understanding of the choices they are expected to make.

Several efforts have researched interfaces intended to improve the way in which users are informed about mobile app data collection and usage practices. Kelley et al. evaluated the benefits of including privacy facts in an app's description in the app store, effectively enabling users to take into account privacy considerations prior to download time [7]. Choe et al. showed that a framing effect can be exploited to nudge people away from privacy invasive apps [37]. The National Telecommunications and Information Administration (NTIA) released guidelines for a short-form mobile app privacy notice in July 2013, aiming to provide app users with clear information about how their personal data are collected, used and shared by apps [38, 39]. Work by Balebako et al. [40], suggests that more work may be required for these interfaces to become truly effective. More generally, Felt et al. discussed the strengths and weaknesses of several permission-granting mechanisms and provided guidelines for using each mechanism [41].

Studies have also shown that users are often surprised when they find out about the ways in which information collected by their apps is being used [13, 42, 43], e.g. what type of data is requested, how often, and for what purpose. In [13], we used crowdsourcing to identify app-permission-purpose triples that were inconsistent with what users expected different apps to collect. We further showed that such deviations are often closely related with lack of comfort granting associated permissions to an app. Our paper builds on this earlier work by scaling up our crowdsourcing framework and performing more advanced data analysis to allow for the development of finer privacy preference models. Our main contribution here is not only to show how mobile app privacy preferences vary with the purpose of app permission pairs but also in the form of a taxonomy of purposes, which we can later leverage to identify clusters of like-minded users.

2.3 Privacy Preference Learning

A first data mining study of mobile app permissions was presented by Frank et al., where they authors looked for permission request patterns in Android apps [44]. Using matrix factorization techniques, they identified over 30 common patterns of permission requests. Rather than looking for patterns of

[1] The Android manifest file of each app presents essential information about this app to the Android system, information the system must have before it can run any of the app's code.

permission requests, our work in this area aims to identify patterns in user privacy preferences, namely in the willingness of users to grant permissions to mobile apps for different purposes.

This work more closely aligned with an earlier study published by three of the co-authors, looking at patterns among the Android permission settings of 239,000 LBE Privacy Guard [31] users for around 12,000 apps [12]. In this earlier work, the three co-authors showed that it was possible to define a small number of privacy profiles that collectively captured many of the users' privacy settings. It further explored mixed initiative models that combine machine learning to predict user permission settings with user prompts when the level of confidence associated with certain predictions appears too low. In contrast to analyzing actual user privacy settings, our work focuses on deeper privacy models, where we elicit people's privacy preferences in a context where they are not just about the permissions requested by an app but also about the one or more purposes associated with these requests (e.g. to enable the app's core functionality versus to share data with an advertising network or an analytics company). While our results bear some similarity with those presented in [12], they are significant because: (i) they show that the purpose for which an app requests a certain permission has a major impact on people's willingness to grant that permission., and (ii) using these more detailed preference models elicited from better-informed users, it is possible to derive a small number of privacy profiles with significant predictive power.

To the best of our knowledge, our work on quantifying mobile app privacy preferences is the first of its kind. It has been influenced by earlier work by several of the co-authors on building somewhat similar models in the context of user location privacy preferences. [45-52]. For example, Lin et al. [45] suggested that people's location-sharing privacy preferences, though complicated, can still be modeled quantitatively. Early work by Sadeh et al. [52] showed that it was possible to predict people's location sharing privacy preferences and work by Benisch et al. explored the complexity of people's location privacy preferences [51]The work by Ravichandran et al. [46] suggested that providing users with a small number of canonical default policies can help reduce user burden when it comes to customizing the fine-grained privacy settings. The work by Cranshaw et al. [47] applied a classifier based on multivariate Gaussian mixtures to incrementally learn users' location sharing privacy preferences. Kelley et al [49] and later Mugan et al. [48] also introduced the notion of understandable learning into privacy research. They used default personas and incremental suggestions to learn users' location privacy rules, resulting in a significant reduction of user burden. Their results were later evaluated by Wilson et al. [50] in a location sharing user study.

As pointed out by Wilson et al. with regard to location sharing privacy in [50], "… the complexity and diversity of people's privacy preferences creates a major tension between privacy and usability…" The present mobile app privacy research is motivated by a similar dilemma, which extends well beyond just location. It shows that approaches that worked well in the context of location sharing appear to offer similar promise in the broader context of mobile app privacy preferences, with a methodology enhanced with the use of static analysis to identify the purpose of mobile app permissions.

3. DATA COLLECTION

Before analyzing people's privacy preferences of mobile apps, it is necessary to gain a deeper understanding of mobile apps with regard to their privacy-related behaviors as well as the implication of these behaviors. In this section, we provide technical details of how we leveraged static analysis to dissect apps and what we learnt.

3.1 Downloading Android Apps and Their Meta-data

We crawled the Google Play web pages in July 2012 to create an index of all the 171,493 apps that were visible to the US users, among which 108,246 of them were free apps. We obtained the metadata of these apps, including the app name, developer name, ratings, number of downloads, etc. We also downloaded all the binary files of free apps through an open-source Google Play API [3]. Note that Google has strict restrictions on app purchase frequency and limits the number of apps that can be purchased with a single credit card. Because of these restrictions, we opted to only download and analyze free apps in this work. Additional analysis using similar method of our work can be applied to paid apps as well.

3.2 Analyzing Apps' Privacy-Related Behaviors

We used static analysis tools given that they are more efficient and easier to automate. We chose Androguard [53] as our major static analysis instrument. Androguard is a Python based tool to decompile Android apk files and to facilitate code analysis. We focused our analysis on the top 11 most sensitive and frequently used permission as identified earlier [19]. They are: INTERNET, READ_PHONE_STATES, ACCESS_COARSE_LOCATION, ACCESS_FINE_LOCATION, CAMERA, GET_ACCOUNTS, SEND_SMS, READ_SMS, RECORD_AUDIO, BLUE_TOOTH and READ_CONTACT. We created our own analysis scripts with the Androguard APIs and identified the following information related to apps' privacy-related behaviors: 1) permission(s) used by each app; 2) The classes and segments of code involved in the use of permissions; 3) All the 3rd-party libraries included in the app; 4) Permissions required by each 3rd-party library. The analysis of 3rd-party libraries provided us more semantic information of how users' sensitive data were used and to whom they were shared.

We obtained permission information of each app by parsing the manifest file of each apk file. We further scanned the entire de-compiled source code and looked for specific Android API calls to determine the classes and functions involved in using these permissions. We identified 3rd-party libraries by looking up package structures in the de-compiled source code. It is possible that we may have missed a few libraries, though we are pretty confident that we were able to correctly identify the vast majority of them and in particular the most popular ones. For the sake of simplicity, we did not distinguish between different versions of the same third party library in our analysis. Similar to the permission analysis step described above, the permission usage of each 3rd-party library was determined by scanning through all the Android standard API calls that relate to the target permission in the de-compiled version of the library's source code.

We further leveraged five Amazon EC2 M1 Standard Large Linux instances to speed up our analysis of this large quantity of

Table 1. Nine categories of 3rd-party libraries

Type	Examples	Description
Utility	Xmlparser, hamcrest	Utility java libraries, such as parser, sql connectors, etc
Targeted Ads	admob, adwhirl,	Provided by mobile behavioral ads company to display in-app advertisements
Customized UI Components	Easymock, kankan,	Customized Android UI components that can be inserted into apps.
Content Host	Youtube, Flickr	Provided by content providers to deliver relevant image, video or audio content to mobile devices.
Game Engine	Badlogic, cocos2dx	Game engines which provide software framework for developing mobile games.
SNS	Facebook, twitter,	SDKs/ APIs to enable sharing app related content on SNSs.
Mobile Analytics	Flurry, localytics	Provided by analytics company to collect market analysis data for developers.
Secondary Market	Gfan, ximad, getjar…	Libraries provided by other unofficial Android market to attract users.
Payment	Fortumo, paypal, zong…	e-payment libraries

apps. The total analysis required 2035 instance hours, i.e. approximately 1.23 minutes per app. Among all the 108,246 free apps, 89,903 of them were successfully decompiled (83.05%). Upon manual inspection of a few failure examples, we observed that failure to de-compile was primarily attributed to code obfuscation.

In the static analysis, we identified over a thousand 3rd-party libraries used by various apps. We looked up the top 400 3rd-party libraries that are most frequently used in all these apps to understand the purpose or functionality associated with each, based on which we organized these 3rd-party libraries into 9 categories as detailed in Table 1[2]. These categories include Targeted Advertising, Customized UI Components, Content Host, Game Engine, Social Network Sites (SNS), Mobile Analytics, Secondary Market, Payment and other Utilities. We also analyzed how different types of resources (permissions) were used for various purposes. For all the apps we analyzed, we observed an average usage of 1.59 ($\sigma = 2.82$, median=1) 3rd-party libraries in each app. There were some extreme cases where an app used more than 30 3rd-party APIs. For example, the app with the package name "com.wikilibs.fan_tatoo_design_for _women_2" used 31 3rd-party libraries, 22 of which were targeted advertising libraries, such as adwhirl, mdotm, millenialmedia, tapjoy, etc. In the majority of cases (91.7%), apps are bundled with less than or equal to 5 different 3rd-party libraries. The targeted advertising libraries are found in more than 40% of these apps. SNS libraries achieved an average penetration of 11.2% of the app market, and mobile analytics libraries had an average penetration of 9.8% of the app market.

In additional to these nine categories of sensitive data uses by third parties, we also used "internal use" to label sensitive data usages caused by the application itself rather than a library. It should be noted that, for these internal uses, we currently cannot determine why a certain resource is used (e.g., whether it is "for navigation", "for setting up a ringtone", etc.). Based on existing practices, the fact that the API call is within the app's code rather than in a 3rd party library indicates a high probability that the resource is accessed because it is required by the mobile app itself rather than to collect data on behalf of a third party.

Our static analysis provided a systems-oriented foundation for us to better understand mobile apps in terms of their privacy-related behaviors, which enabled us to study users' preferences in regard to these app behaviors in the later part of this paper. Note that, although we only collected users' preferences of 837 apps among the apps we dissected as described in the following subsection, the static analysis of 89,000 + apps was necessary for us to understand the bigger picture of sensitive data uses and to identify the nine categories of 3rd-party libraries.

3.3 Crowdsourcing Users' Mobile App Privacy Preferences

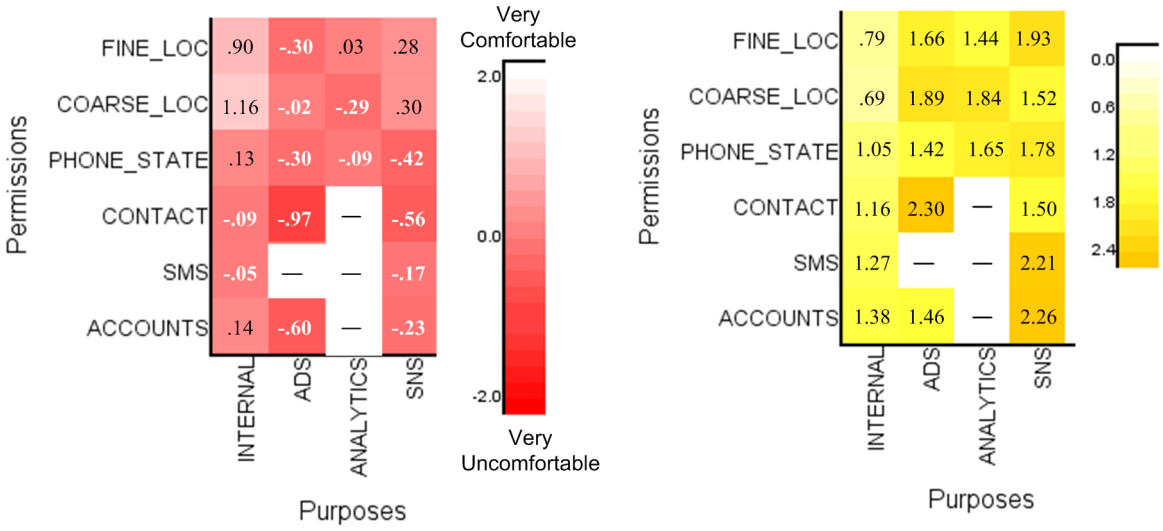
To link users' privacy preferences to these app behaviors we identified through static analysis, we leveraged Amazon Mechanical Turk (AMT) to collect users' subjective responses through a study similar what Lin et al. did in [13]. Participants were shown the app's icon, screen shots, and description of apps. Participants were asked if they expected this app to access certain type of private information and were also asked how comfortable (from "-2" very uncomfortable to "+2" very comfortable) they felt downloading this app given the knowledge that this app accesses their information for the given purposes. Each HIT (Human Intelligence Task) examined one app – permission – purpose triple that we identified as described in the previous section. For example, in one HIT, participants were asked to express their level of comfort in letting Angry Birds (app) access their precise location (permission) for delivering targeted ads (purpose). We added one qualification question in each HIT, asking participants to select from a list of three app categories, to test whether they had read the app's description and whether they were paying attention to the questions. The template of the HIT is shown in Appendix A.

In total we published 1200 HITs on AMT, probing 837 mobile apps that we randomly sampled from the top 5000 most popular free apps. For each HIT, we aimed to recruit 20 unique

Table 2. Participants' demographic summary

Education	**%**	**Age Group**	**%**
High School	31%	Under 21	11%
Bachelor Degree	63%	21-35	69%
Graduate Degree	6%	36-50	16%
		51-65	3%
Gender	**%**	Over 65	1%
Female	41%		
Male	59%		

[2] The library uses follows a power-law distribution, therefore, the top 400 most popular libraries covered over 90% of uses.



(a) Average user preferences

(b) Variances in user preferences

Figure 1 (a) The average self-reported comfort ratings of different permission usages. The lighter shades represent permission-purpose pairs users are more comfortable granting, whereas the darker shades of red indicate less comfort. (b) The variances in comfort levels. Many entries have large variances. Entries with a short dash indicate the absence of data for a particular permission-purpose.

participants to answer our questions. Participants were paid $0.15 per HIT. We restricted our participants to U.S. smartphone users with previous HIT approval rate higher than 90%.

The study ran for 3 weeks starting on June 15th, 2013. After the data collection period, we first eliminated responses that failed the qualification questions (~7%), and then we eliminated 39 HITs because they had less than 15 responses. This yielded a dataset of 21,657 responses contributed by 725 AMT workers.

4. DESCRIPTIVE RESULTS

4.1 Participants

We collected demographic information of our participants including gender, age and education background to help us analyze our data, though we did not specifically control the gender ratio or any other demographic composition of our participants. Among these participants, 41% of them were female; 69% of participants were between 21 and 35, 16% of them are between 36 and 50 (see Table 2). We also observed that more than 60% of the participants were reported to have a bachelor's degree or equivalent and 6% had a master's degree or PhD. The average education level of our participants was significantly higher than the average education level of the entire U.S. population as reported in [54]. Compared to the demographics of crowd workers as reported in [55], our participant pool contains more people with bachelor's degrees and fewer with graduate degrees.

This difference in demographics may be caused by self-selection, since usually crowd workers would be more likely to work on HITs that interest them. However, other data collection methods, such as Internet surveys, often have similar sampling problems. While this sample bias has to be taken into account when interpreting our results, we suspect that our study is no worse than most others in terms of the representativeness of our participant pool.

4.2 Users' Average Preferences and Their Variances

To visualize our results, we aggregated self-reported comfort ratings by permission and purpose. Figure 1 (a) shows the average preferences of all 725 participants, where white indicates participants were very comfortable (2.0) with the disclosure, and red indicates very uncomfortable (-2.0). In other words, darker shades of red indicate a higher level of concern. Entries with a short dash indicate the absence of data for a particular permission-purpose. For example, in our analysis, we did not see any analytics library accessing users' contact information or trying to send or receive SMS. Note that these heat map visualizations only display the most important six permissions and four purposes, since they are the most popular data uses and the sources of the primary distinctions among users (which we will introduce in the next subsection).

The three use cases with the highest levels of comfort were: (1) apps using location information for their internal functionality (fine location: $\mu = 0.90$, coarse location: $\mu = 1.16$); (2) SNS libraries bundled in mobile apps using users' location information so this context information can be used in sharing (fine location: $\mu = 0.28$, coarse location: $\mu = 0.30$); (3) apps accessing smartphone states, including unique phone IDs, and account information for internal functionality ($\mu = 0.13$).

For the remaining cases, users expressed different levels of concerns. Users were generally uneasy with (1) targeted advertising libraries accessing their private information, especially for their contact list ($\mu = -0.97$) and account

information[3] (μ = -0.60); (2) SNS libraries that access their unique unique phone ID (μ = -0.42), contact list (μ = -0.56), as well as information related to their communication and web activities such as SMS (μ = -0.17) and accounts (μ = -0.23); and (3) mobile analytic libraries accessing their location (μ = -0.29) and phone state[4] (μ = -0.09).

This aggregation of data gave us a good starting point to spot general trends in users' privacy preferences. At the same time, these are averages and, as such, they do not tell us much about the diversity of opinions people might have. An important lesson we learnt from previous literature of location privacy is that users' privacy preferences are very diverse. To underscore this point, we plotted the variances of user preferences of the same use cases, as shown in Figure 1 (b). Here, darker shades of yellow indicate higher variance among users' comfort rating for different purposes.

Figure 1 (b) shows that users' preferences are definitely not unified. Variances are larger than 0.6 (of a rating in a [-2, +2] scale) in all cases. In 25% of cases, variances exceeded 1.8. Users' disagreements were highest in the following cases, including: (1) SNS libraries accessing users' SMS information as well as their accounts; (2) targeted advertising libraries accessing users' contact list; (3) users' location information being accessed by all kinds of external libraries.

This high variance in users' privacy preferences suggests that having a single one-size-fits-all privacy setting for everyone may not work well – at least for those settings with a high variance. We cannot simply average the crowdsourced user preferences and use them as default settings as suggested in [32]. This begs the question of whether users could possibly be subdivided into a small number of groups or clusters of like-minded individuals for which such default settings (different settings in different groups) could be identified. We discuss this idea in the next section.

5. LEARNING MOBILE APP PRIVACY PREFERENCES

Given the large variances identified above, a unified default setting evidently cannot satisfy all the users' privacy preferences. Therefore, we chose to investigate methods for segmenting the entire user population into a number of subgroups that have similar preferences within the subgroups. Then by identifying the suitable default settings for each of these groups and the group each user belongs to, we can suggest individual users with more accurate default settings.

5.1 Pre-processing

To identify these groups, we need to properly encode each user's preferences into a vector and trim the dataset to prevent over-fitting. More specifically, we conducted three kinds of preprocessing before feeding the dataset into various clustering algorithms. First, we eliminated participants who contributed less than 5 responses to our data set, since it would be difficult to categorize participants if we know too little about their preferences. This step yielded a total number of 479 unique participants with 20,825 responses. On average, each participant contributed 43.5 responses (σ = 38.2, Median=52). Second, we aggregated a participant's preferences by averaging their indicated comfort levels of letting apps use specific permissions for specific purposes. "NA" is used if a participant did not have a chance to indicate his/her preferences for a given permission-purpose pair. Lastly, for each missing feature ("NA"), we found the k (k=10) nearest neighbors that had the corresponding feature. We then imputed the missing value by using the average of corresponding values of their neighbor vectors.

After these preprocessing steps, we obtained a matrix of 77 columns (i.e. with regard to 77 permission-purpose pairs) and 479 rows, where each row of the matrix represented a participant. Each entry of the matrix was a value between [-2, +2]. This preference matrix was free of missing values.

5.2 Selection of Algorithms and Models

We opted to use hierarchical clustering with an agglomerative approach to cluster participants' mobile app privacy preferences. In the general case, the time complexity of agglomerative clustering is $O(n^3)$ [56]. Though its time complexity is not as fast as k-means or other flat clustering algorithms, we chose hierarchical clustering mainly because its resulting hierarchical structure is much more informative and more interpretable than unstructured clustering approaches (such as k-means). More specifically, we experimented with several distance measures [56], including Euclidean distance, Manhattan distance [57], Canberra Distance [58], and Binary distance [59]. We also experimented with four agglomerative methods, including Ward's method [60], Centroid Linkage Method [61], Average Linkage method [61], and McQuitty's Similarity method [62].

We limited our exploration to the above-mentioned distance functions and agglomerative methods, since other distance functions or agglomerative methods either produce similar results as the above-mentioned ones or are not appropriate for our tasks based on the characteristics of our data. As research on clustering techniques continues, it is possible that new techniques could provide even better results than the ones we present. We found however these techniques were already sufficient to isolate very different categories of mobile apps, when it comes to their permissions and the purposes associated with these permissions.

To select the best model, we experimented with various ways of combining the four agglomerative methods and four distance measures and also varied the number of clusters k from 2 to 20 by using the R package "hclust" [63]. We conducted all the experiments on a Linux machine which has XeonE5-2643 3.3GHz CPU (16 cores) and 32G memory. We had two selection criteria in determining which combination of distance function and agglomerative method to use. First, the combination should not generate clusters with extremely skewed structures in dendrograms. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. The tree structure in the dendrogram illustrate how clusters merged in each iteration. We check this by heuristically inspecting the dendrograms of each clustering result. The other criteria is the combination of three internal measures, namely connectivity [64], Silhouette Width [65] and Dunn Index [66].

[3] GET_ACCOUNTS permission gives apps the ability to discover existing accounts on managed by Android operating system without knowing the passwords of these accounts.

[4] READ_PHONE_STATE permission gives apps the ability to obtain unique phone id and detect if the users is currently calling someone.

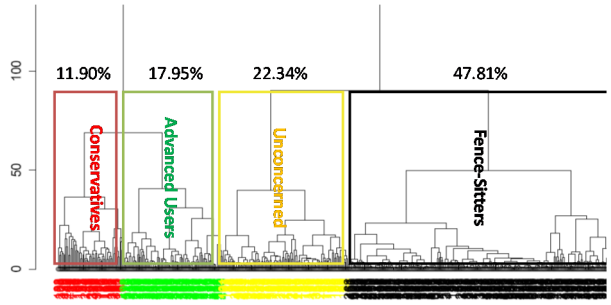


Figure 2. The resulting dendrogram produced by hierarchical clustering with Canberra distance and average linkage agglomerative method. Four different colors are used to indicate the cluster composition when *k=4*. We also overlay the cluster names on the dendrogram which will be explained in Section 6.1.

These three internal measures validate the clustering results based on their connectivity, compactness and degree of separation.

5.3 Resulting Clusters

Based on the two criteria described in the previous sub-section, we obtained the best clusters by using Canberra distance and Average Linkage method with k=4.

Figure 2 illustrates the resulting dendrogram produced by the above-mentioned clustering configurations, where four different colors indicate the four clusters when k=4. Among the four identified clusters, the largest one (colored in black in Figure 2) includes 47.81% of instances, whereas the smallest cluster (colored in red) includes 11.90% instances. We assigned a name to each cluster based on its outstanding characteristics and overlaid these names on the dendrogram as well. The explanation of these names and the interpretation of our clustering results are discussed in the following section.

6. RESULT INTERPRETATION

To make sense of what these clusters mean, we computed the centroid of each cluster by averaging the feature vectors of instances within the cluster. Note that we computed the centroid of each cluster based on the non-imputed data points, i.e. only averaging the entries when there were true values, since they better estimate the true average preferences of users in each category.

6.1 Making Sense of User Clusters

We used a heat map to visualize these clusters[5] as shown in Figure 3 – Figure 6. The vertical dimension of these heat maps represents the uses of different permissions, and the horizontal dimension represents why a certain permission is requested. In each figure, the left grids represent the centroid of the cluster. We use two colors to indicate people's preferences. White indicates that participants feel comfortable with a given permission-purpose whereas shades of red indicate discomfort, with darker shades of red corresponding to greater discomfort. The right grids in each figure show the corresponding variances within the cluster. Compared to the variances in Figure 1, the variance of each clusters are significantly smaller. Some of them are almost negligible.

We have labeled each cluster with a name that attempts to highlight its distinguishing characteristics. The labels are (privacy) "*conservatives*", "*unconcerned*", "*fence-sitters*", and "*advanced users*".

The (Privacy) Conservatives: Although conservatives form the smallest group among the four clusters, they still represent 11.90% of our participants (see Figure 3). Compared to the heat maps of other clusters, this cluster (or "privacy profile") has the largest area covered in red and also the overall darkest shades of red (indicating the lack of comfort granting permissions). In general, these participants felt the least comfortable granting sensitive information and functionality to third parties (e.g., location and unique phone ID). They also felt uncomfortable with mobile apps that want to access their unique phone ID, contacts list or SMS functionality, even if for internal purposes only.

The Unconcerned: This group represents 23.34% of all the participants and forms the second largest cluster in our dataset (Figure 4). The heat map of this privacy profile has the largest area covered in light color (indicate of comfort). In general, participants who share this privacy profile showed a particularly high level of comfort disclosing sensitive personal data under a wide range of conditions, no matter who is collecting their data and for what purpose. The only concerning (red) entry in the heat map is when it comes to granting SNS libraries access to the GET_ACCOUNTS permission (e.g. information connected to accounts such as Google+, Facebook, YouTube). A closer analysis suggests that it might even be an anomaly caused by the lack of sufficient data points for this particular entry. Another possible interpretation might be that a considerable portion of participants did not understand the meaning of this permission and mistakenly thought this permission gives apps ability to know their passwords of all accounts

The Fence-Sitters: We labeled participants in this cluster as "Fence-Sitters" because most of them did not appear to feel strongly one way or the other about many of the use cases (Figure 5). This cluster represents nearly 50% of our population. Unsurprisingly, this group of participants felt quite comfortable letting mobile apps access sensitive personal data for internal functionality purposes. When their information is requested by 3rd-party libraries such as for delivering targeted ads or conducting mobile analytics, their attitude was close to neutral (i.e. neither comfortable nor uncomfortable). This is reflected in the heat map with large portions of it colored in light shades of pink (close to the middle color in the legend). This group of participants also felt consistently comfortable disclosing all types of sensitive personal data to SNS libraries. Further research on why so many participants behave in this way is challenging and necessary. We suspect that this might be related to some level of habituation or warning fatigue, namely they might have gotten used to the idea that this type of information is being accessed by mobile apps and they have not experienced any obvious problem resulting from this practice.

This cluster of participants also reminds us of the privacy pragmatist group identified by Westin in producing privacy

[5] Again, in these visualizations, we only display the most important six permissions and four purposes that strongly differentiate participants.

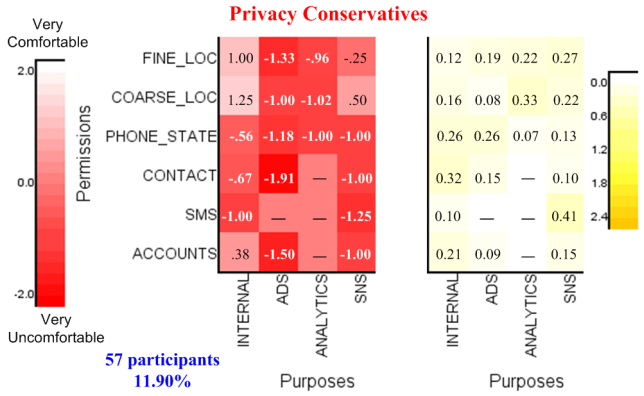


Figure 3. The centroid (left) and variances (right) of Privacy Conservatives. This group of participants expressed the most conservative preferences. They did not like their private resources used by any external parties. Notice how much lower the variances are relative to those in Figure 1.

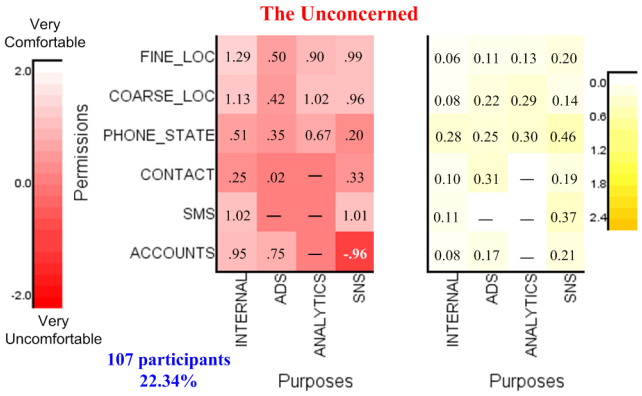


Figure 4. The centroid (left) and variances (right) of the unconcerned. This group of participants felt comfortable disclosing their data to 3rd-parties for most cases.

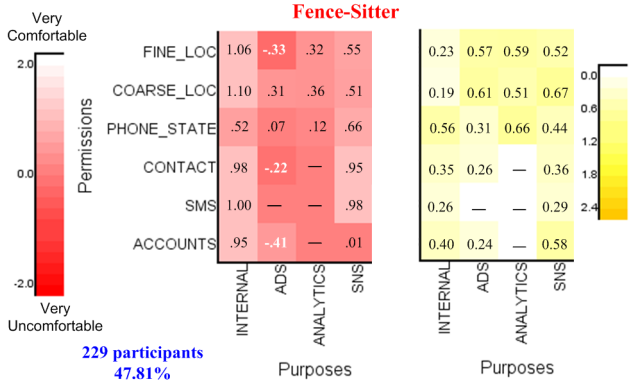


Figure 5. The centroid (left) and variances (right) of the fence-sitters. This is the largest cluster in our study. This group of participants felt neutral to ads and mobile analytics. This group also had the largest within-cluster variances.

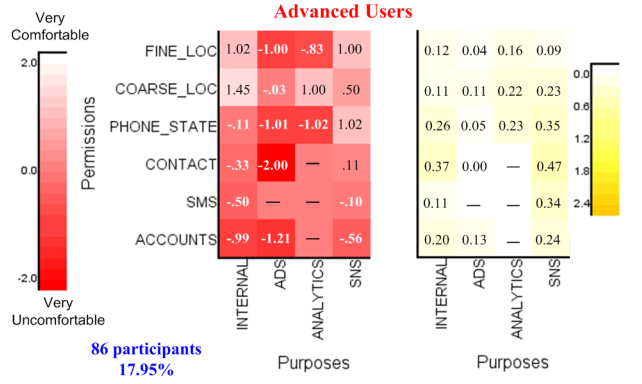


Figure 6. The centroid (left) and variances (right) of advanced users. This group of users were more selective in their privacy preferences.

indexes [67]. Westin found that while small numbers of users would fall at both extremes of the spectrum (i.e. privacy fundamentalist, and unconcerned), the majority of users tend to be in-between (pragmatists). An interesting finding of our analysis is that the preferences of these middle-of-the-road users can generally be captured with just two profiles, namely the "fence-sitters" and the "advanced users" (see next subsection).

The Advanced Users: The advanced user group represents 17.95% of the population (see Figure 6). This group of participants appeared to have a more nuanced understanding of what sorts of usage scenarios they should be concerned about. In general, most of them felt comfortable with their sensitive data being used for internal functionality and by SNS libraries. One possible reason of why they felt okay with the latter scenario is because they still have control over the disclosures, since these SNS libraries often let people confirm sharing before transmitting data to corresponding social network sites. In addition, this group disliked targeted ads and mobile analytic libraries, but still felt generally agreeable to disclosing context information at a coarser level of granularity (i.e. coarse location). This observation again suggests that this group of users have better insight when it comes to assigning privacy risks to different usage scenarios.

6.2 Estimating the Predictive Power of the Clusters

As discussed above, the clusters we have identified give rise to significant drops in variance. Could these or somewhat similar clusters possibly help predict many of the permission settings a user would otherwise have to manually configure? Providing a definite answer to this question is beyond the scope of this paper, in part because our data captures preferences (or comfort levels) rather than actual settings and in part also because answering such a question would ultimately require packaging this functionality in the form of an actual UI and evaluating actual use of the resulting functionality. Below we limit ourselves to an initial analysis, which suggests that the clusters we have identified have promising predictive power and that similar clusters could likely be developed to actually predict many permission settings – for instance in the form of recommendations.

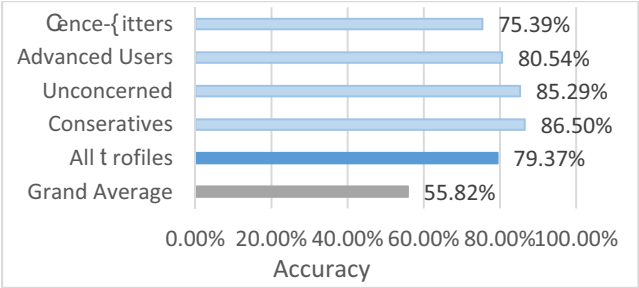


Figure 7. Compared to using a single one-size-fits-all grand average profile to all participants, classifying participants into four profiles can significantly increase the accuracy in predicting if the system should grant , deny or prompt users for a specific app-permission-purpose triple (55.82% vs. 79.37%). For two profiles ("unconcerned" and "conservatives") the prediction accuracies are higher than 85%. All numbers were averaged over 10 runs with different partitions of training and testing data.

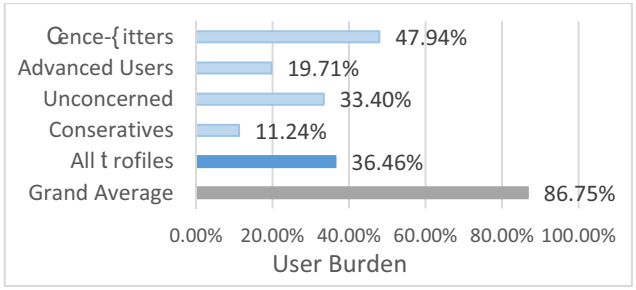


Figure 8. Choosing a good privacy profile reduces the user configuration effort down to just 36.5% of all app-permission-purpose triples, whereas users would need to configure nearly 87% of the triples if one were to rely on a single one-size-fits-all grand profile. For users in the "advanced" and "conservative" categories, user burden drops below 20%. All numbers were averaged over 10 runs using different partitions of training and testing data and were weighted by the usages of all permission-purpose pairs among the 837 apps.

Specifically, as part of our analysis, we transformed the four cluster centroids into four "*privacy profiles*" (i.e. sets of recommendations) by quantizing the [-2, 2] comfort rating into three options, namely "Accept" (average comfort rating higher than or equal to 0.67), "Reject" (average comfort rating lower than or equal to -0.67), and "Prompt" (average comfort rating between -0.67 and +0.67 exclusively). In other words, in our analysis, we assumed that "Accept" meant the corresponding purpose-permission pair would be automatically granted. Similarly a "Reject" value is interpreted as automatically denying the corresponding permission-purpose pair. Cases with values falling in between are simply assumed to result in a user prompt, namely asking the user to decide whether to grant or deny the corresponding permission-purpose pair. In short, under these assumptions, a user would be assigned a profile, which in turn would be used to automatically configure those permission-purpose settings for which the profile has an "Accept" or "Reject" entry, with the remaining settings having to be manually configured by each individual user.

We now turn to our estimation of the potential benefits that could be derived from using clusters and privacy profiles to help users configure many of their app-permission-purpose settings. The results presented here are based on assumptions made about how one could possibly interpret the preferences we collected and treat them as proxies for actual settings users would want to have. While we acknowledge that an analysis under these assumptions is not equivalent to one based on actual settings and that the clusters and profiles one would likely derive from actual settings would likely be somewhat different, we believe that the results summarized below show promise both in terms of potential predictive power and potential reductions in user burden.

We randomly split all the participants into 10 folds of (almost) identical sizes. We then used each possible combination of 9 folds of participants to compute cluster centroids and generate privacy profiles (in terms of "Accept", "Deny", and "Prompt" for each permission-purpose pair). The remaining fold of participants was used to evaluate the benefits of the learned profiles – both in terms of expected increase in accuracy and in terms of expected reductions in user burden. We assumed that all testing participants were able to choose a privacy profiles that closely captured their preferences (which will be discussed in Subsection 6.3-6.4). We averaged the following two metrics across all 10 runs:

(1) **Accuracy**: the percentage of time that the selected privacy profile agreed with the comfort rating provided by each individual participants in the testing group for each of the app-permission-purpose triples available in the data set for that user. (Figure 7).

(2) **User burden**: the percentage of time the participants in testing sets would be prompted to specify their decisions, weighted by the usages of all permission-purpose pairs among all apps (Figure 8). These usages were measured by calculating the percentage of apps in crowdsourcing study (837 in total) that use a specific permission for a specific purpose.

To evaluate the benefits of the profiles, we compare both of these metrics, as obtained using our profiles, with identical metrics obtained using a single one-size-fits-all grand profile for all users (as shown in Fig. 1 (a)). This is referred to as "Grand average profile".

As can be seen in Figure 7, the profiles result in an overall accuracy of nearly 80% (79.37%). In comparison predictions based on a single one-size-fits-all model result in an accuracy of merely 56%, which is not much better than simply prompting users all the time. In particular, using our four profiles, accuracies for people falling in the "unconcerned" and "conservative" groups are higher than 85%.

Figure 8 shows how under our assumptions applying privacy profiles as default settings could significantly reduce user burden. In particular, when using a single- one-size-fits-all model, users would on average have to be prompted for nearly 87% of all their app-permission-purpose triples. In contrast, when using the four privacy profiles, the number of prompts drops to 36.5% of the user's total number of app-permission-purpose triples. This clearly represents a significant reduction in user burden. For users falling in the "advanced" and "conservative" categories the number of prompts drops below 20%. While we acknowledge that further research is required, using actual permission settings

rather than measures of comfort levels, we believe that the results of our analysis show great promise and warrant further work in this area.

6.3 Do Demographics Matter?

Now we want to see how to assign users to the privacy profiles that most closely capture their privacy preferences. Here we first look at whether users' demographic information – including gender, age and education level – is sufficient to determine which privacy profile a user should be assigned. This included looking at the distribution of gender, age and education level in each user cluster and also looking at variance (ANOVA) to see if there are significant differences in these distributions.

In general, we found that in regard to the gender distribution, a one-way analysis of variance yield NO significant differences between groups, *F(3, 475)=2.049, p=0.106*. For age distribution, we encoded the age groups as (1:= under 21, 2:= age 21-35, 3:=age 36-50, 4:=age 51-65, 5:=above 65) in our calculation. A one-way analysis of variance reveals **significant differences between groups in regard to age distribution**, *F(3, 475)=4.598, p=0.003*. Post hoc analyses also reveals that the unconcerned group on average are younger ($\mu = 1.69$, $\sigma = 0.57$) than other groups combined ($\mu = 1.91$, $\sigma = 0.76$), and the advanced user group on average are older ($\mu = 2.05$, $\sigma = 0.61$) than other groups combined ($\mu = 1.83$, $\sigma = 0.71$).

We also performed a similar test on the education level of all four groups of participants. We encoded the education levels such that "1" stands for high school or lower level of education, "2" stands for bachelor or equivalent level of degrees, and "3" stands for master's or higher level of degrees. An ANOVA test shows that **the effect of education level was strongly significant**, *F(3, 475)=7.52, p=6.3E-05*. Post hoc analyses show that the conservatives ($\mu = 1.65$, $\sigma = 0.48$) and the unconcerned ($\mu = 1.67$, $\sigma = 0.54$) have lower education levels compared to the remaining groups combined ($\mu = 1.85$, $\sigma = 0.57$), and the advanced users ($\mu = 2.01$, $\sigma = 0.60$) are more likely to have a higher level of education.

Although there are statistically significant effects in demographics, a regression from demographic information to the cluster label yields accuracy no better than directly putting every user as Fence-Sitters. In other words, we should not directly use gender, age, or education level to infer which privacy profile should be applied to individual user. This does not mean however that in combination with other factors, these attributes would not be useful. Below, we seek more deterministic methods to assign privacy profiles in the following sub-section.

6.4 Possible Ways to Assign Privacy Profiles

We start with a typical scenario where a privacy profile can be assigned to a user. When a user boots up her Android device for the first time (or possibly at a later time), the operating system could walk her through a "wizard" and determine which privacy profile is the best match for her. The profile could then be used to select default privacy settings for this user. As the user downloads apps on the smartphone, "App Ops" or some equivalent functionality would then be able to automatically infer good default settings for the user. The major challenge here is how we can accurately determine which cluster this user belongs to without any previous data about this user.

One possible way is to ask users to label a set of mobile apps. We could present users with a small set of example apps together with detailed descriptions such as the sensitive data collected by these apps and for what purposes. Users could rate each app based on its sensitive data usages. We could then classify users based on these ratings. This would work well if we could identify a small number of particularly popular apps that can differentiate between users - say just asking people whether they feel comfortable sharing their location with Angry Birds game for advertising purpose and whether they feel comfortable posting their location on Facebook through the Scope app. Further research on selecting the most effective set of apps would make this process more effective and stable.

Alternatively, we might probe users' privacy preferences by asking them a small set of general questions. Similar ideas have been suggested for helping users set up their location sharing rules [46] [48]. In particular Wilson et al. in [50] described a simple wizard for the Locaccino system, where a small number of questions were asked to guide users through the selection of good default location sharing profiles. A similar method could be used to identify a small number of questions to help determine appropriate mobile app privacy profiles for individual users.

Given the four privacy profiles that we identified, we note several observations that could be used to differentiate between different groups of users. For example, the reported comfort ratings with respect to sharing data with advertising agencies can be used to separate the unconcerned group from the privacy conservatives and the advanced users; we could use people's preferences with regard to sharing coarse location information for mobile analytics to further differentiate between the latter two groups; or we can isolate the privacy conservatives based on their extreme negative comfort rating with SNS libraries. One should be able to identify a small number of questions based on these or similar observations. The ideal scenario would be that, based on their answers to these questions, users could be accurately assigned to the most appropriate cluster. For example, we can ask one question with regard to targeted advertising, such as "How do you feel letting mobile apps access your personal data for delivering targeted ads?" or questions about mobile analytics, such as "How do you feel about letting mobile apps share your approximate location with analytics companies?" The exact wording and expressions used in these questions would obviously need to be refined based on user studies.

The privacy profiles we extracted are a good estimation but might not perfectly match individual user preferences. It is necessary to clarify that applying privacy profiles does not prevent users from further personalizing their privacy decisions. In addition to choosing an appropriate privacy profile as a starting point, users could be provided with user-oriented machine learning functionality or just interactive functionality that helps them iteratively refine their settings [47-49].

7. DISCUSSION

7.1 Limitations of This Work

This work has several limitations. For example, our study focused solely on free apps downloaded from the Google Play. Apps that require purchase might exhibit slightly different privacy-related behaviors with regard to what sensitive resources to request and for what purpose. There are two major challenges that prevented us to investigate paid apps: (a) the monetary cost of purchasing a large number of paid apps would be substantial (we estimate over $80K to get all the paid apps); (b) there is no way to programmatically do batch purchasing on Google Play, since

Google limits the frequency of app purchases using a single credit card in a single day. It should also be noted that free apps represent the majority of app downloads, and paid apps tend to request fewer permissions – in other words, they give rise to a somewhat smaller number of privacy decisions. This being said, there is no reason to believe that the models derived for free apps could not be extended to paid apps – while people's privacy preferences might be different, there is no reason to believe that similar clusters could not be identified.

In determining why certain sensitive resources are requested, our study used a relatively coarse classification. Our static analysis cannot give finer-grained explanations, such as requesting location for navigation vs. requesting location for nearby search. We acknowledge that our approach is not perfect. However, comparing to a finer analysis relying on manual inspection, using libraries to infer the purpose of permissions enables us to conduct our analysis at large scale. Additional techniques could possibly be developed over time to further increase accuracy. For example, the tool described by Amini et al. [26] that combines crowdsourcing and dynamic analysis might be able to provide this level of details, through it has not been publicly available yet.

Among all the four clusters we identified, the Fence-Sitter cluster has a relatively high variance. By using more advanced clustering techniques better clusters could likely be generated with even smaller intra-cluster variances. However, we consider the primary contribution of this work is to demonstrate the feasibility of profile-based privacy settings. As part of future work, we hope to extend our data collection and experiments, such that we can further refine our clusters and possibly obtain even better results.

7.2 Lessons Learned and Future Prospects

Users' mobile app privacy preferences are not unified. This paper quantitatively proved that mobile app users have diverse privacy preferences. This suggested that simply crowdsourcing people's average preferences as suggested by Agarwal and Hall in the PMP privacy settings [32] might not be optimal. In spite of the diversity, we also show that there are a relatively small number of groups of like-minded users that share many common preferences. Using these identified groups, we derived mobile app privacy preferences profiles, find for each user a profile that is a close match, and use this information to automate the privacy setting process.

Purpose is more important. Previous work in mobile app analysis as well as on users' privacy concerns focused more on identifying the what sensitive information is accessed by apps [17, 42] as well as how often sensitive information is shared with external entities [43]. Lin et al. [13] pointed out the purpose of why sensitive resources are used is important for users to make privacy decision, though they did not quantitative backup this statement. Our work provides crucial evidence to support this statement. The clusters we identified in our participants are more differentiated in the dimension of why these resources are accessed. This finding also provides important implications to privacy interface design in the sense that properly informing users the purposes of information disclosures are at least as important as informing them what information is disclosed. Unfortunately, the current privacy interfaces, such as the Google Play's permission list, fall short in making good explanation of the purposes. We strongly suggest mobile app market owners to consider notifying this important information to their customers.

Make use of the naturally crowdsourced data. In our study, we use Amazon Mechanical Turk as the major platform to collect users' privacy preferences. In reality, given the availability of "App Ops" in Android 4.3, "ProtectMyPrivacy" on jailbroken iPhone, or other similar extensions in rooted Android devices, the operating system or the third-party privacy managers could naturally crowdsource users' privacy preferences without extra effort. These valuable datasets also presumably have better user coverage and are more representative than what we can collect with the limited resources we have. A significant portion of the methodologies discussed in this work can be directly applied to these dataset to build models of mobile users in the wild. We encourage industry to make fully uses of the findings we present in this paper to make real impact in providing users with better privacy controls.

In short, the findings that we present provide important lessons about mobile app users, and also point out a way to make privacy settings potentially usable to end users. However, there is still much work that needs to be done to model users' privacy preferences. We are also aware that users' privacy preferences might keep on evolving and are influenced by the introduction of new technologies and the habituation effect that formed through interacting with the same practices for a long time. Therefore, in addition to all the techniques we proposed, we believe other prospects such as proper user education, improving and enforcing laws and regulations are also crucial and need to be promoted in the long run.

8. CONCLUSION

This paper complements existing mobile app privacy research by quantitatively linking apps' privacy related behaviors to users' privacy preferences. We utilized the static analysis with specific focus on how and why 3rd-party libraries use different sensitive resources and leveraged crowdsourcing to collect privacy preferences of over 700 participants with regard to over 800 apps. Based on the collected data, we identified four distinct privacy profiles, providing reasonable default settings to help users configure their privacy settings. Initial results intended to estimate the benefits of these profiles suggest that they could probably be used to significantly alleviate user burden, by helping predict many of a user's mobile app privacy preferences. Under our proposed approach, users would still be prompted when the variance of the predictions associated with an entry in a given profile exceeds a certain threshold. More sophisticated learning techniques could possibly further boost the accuracy of such predictions.

9. ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation under grants CNS-1012763, CNS-1330596, CNS-1228813 and CNS-0905562, by CyLab under grants DAAD19-02-1-0389, by Army Research Office under W911NF-09-1-0273, by Carnegie Mellon Portugal ICTI 1030348, and by Google.

10. REFERENCES

[1] Wikipedia *App Store (iOS)*. Available: http://en.wikipedia.org/wiki/App_Store_(iOS)

[2] Wikipedia *Google Play*. Available: http://en.wikipedia.org/wiki/Google_Play

[3] Burguera, I., Zurutuza, U. and Nadjm-Tehrani, S. Crowdroid: behavior-based malware detection system for Android. In *Proc. of the SPSM*, 2011.

[4] Felt, A. P., Finifter, M., Chin, E., Hanna, S. and Wagner, D. A survey of mobile malware in the wild. In *Proc. of the SPSM*, 2011.

[5] Kelley, P. G., Consolvo, S., Cranor, L. F., Jung, J., Sadeh, N. and Wetherall, D. A Conundrum of permissions: Installing Applications on an Android Smartphone. In *Proc. of the USEC*, 2012.

[6] Felt, A. P., Ha, E., Egelman, S., Haney, A., Chin, E. and Wagner, D. Android Permissions: User Attention, Comprehension, and Behavior. In *Proc. of the Soups*, 2012.

[7] Kelley, P. G., Cranor, L. F. and Sadeh, N. Privacy as part of the app decision-making process. In *Proc. of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013.

[8] Beresford, A., Rice, A. and Sohan, N. MockDroid: trading privacy for application functionality on smartphones. In *Proc. of the HotMobile*, 2011.

[9] Hornyack, P., Han, S., Jung, J., Schechter, S. and Wetherall, D. These aren't the droids you're looking for: retrofitting android to protect data from imperious applications. In *Proc. of the CCS*, 2011.

[10] Zhou, Y., Zhang, X., Jiang, X. and Freech, V. W. Taming Information-Stealing Smartphone Applications (on Android). In *Proc. of the TRUST*, 2011.

[11] Lunden, I. *U.S. Consumers Avg App Downloads Up 28% To 41; 4 Of 5 Most Popular Belong To Google*. Available: http://techcrunch.com/2012/05/16/nielsen-u-s-consumers-app-downloads-up-28-to-41-4-of-the-5-most-popular-still-belong-to-google/

[12] Liu, B., Lin, J. and Sadeh, N. Reconciling Mobile App Privacy and Usability on Smartphones: Could User Privacy Profiles Help? In *Proc. of the WWW'14*, 2014.

[13] Lin, J., Amini, S., Hong, J., Sadeh, N., Lindqvist, J. and Joy Zhang. Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing. In *Proc. of the Ubicomp'12*, 2012.

[14] Enck, W. *Defending Users against Smartphone Apps: Techniques and Future Directions*. City, 2011.

[15] Enck, W., Ongtang, M. and McDaniel, P. On lightweight mobile phone application certification. In *Proc. of the CCS*, 2009.

[16] Enck, W., Octeau, D., McDaniel, P. and Chaudhuri, S. A Study of Android Application Security. In *Proc. of the USENIX Security Symposium*, 2011.

[17] Enck, W., Gilbert, P., Chun, B.-G., Cox, L., Jung, J., McDaniel, P. and Sheth, A. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones. In *Proc. of the OSDI* 2010.

[18] Chin, E., Felt, A. P., Greenwood, K. and Wagner, D. Analyzing inter-application communication in Android. In *Proc. of the MobiSys*, 2011.

[19] Felt, A. P., Chin, E., Hanna, S., Song, D. and Wagner, D. Android permissions demystified. In *Proc. of the CCS*, 2011.

[20] Felt, A. P., Greenwood, K. and Wagner, D. The effectiveness of application permissions. In *Proc. of the USENIX conference on Web application development*, 2011.

[21] Felt, A. P., Wang, H. J., Moshchuk, A., Hanna, S. and Chin, E. Permission re-delegation: attacks and defenses. In *Proc. of the USENIX conference on Security*, 2011.

[22] Vidas, T., Christin, N. and Cranor, L. Curbing android permission creep. *Proceedings of the Web*, vol. 2, 2011.

[23] *App Profiles*. Available: https://play.google.com/store/apps/details?id=com.appdescriber&feature=search_result#?t=W251bGwsMSwxLDEsImNvbS5hcHBkZXNjcmliZXIiXQ..

[24] Thurm, S. and Kane, Y. I. Your Apps are Watching You. *WSJ*, 2011.

[25] Barrera, D., Kayacik, H. G., Oorschot, P. C. v. and Somayaji, A. A methodology for empirical analysis of permission-based security models and its application to android. In *Proc. of the CCS*, 2010.

[26] Amini, S., Lin, J., Hong, J., Lindqvist, J. and Zhang, J. Mobile Application Evaluation Using Automation and Crowdsourcing. In *Proc. of the PETools*, 2013.

[27] Chin, E., Felt, A. P., Sekar, V. and Wagner, D. Measuring user confidence in smartphone security and privacy. In *Proc. of the Proceedings of the Eighth Symposium on Usable Privacy and Security*, 2012.

[28] Felt, A. P., Egelman, S. and Wagner, D. I've Got 99 Problems, But Vibration Ain't One: A Survey of Smartphone Users' Concerns. In *Proc. of the SPSM*, 2012.

[29] Verduzco, W. *App Ops Brings Granular Permissions Control to Android 4.3*. Available: http://www.xda-developers.com/android/app-ops-brings-granular-permissions-control-to-android-4-3/

[30] Amadeo, R. *App Ops: Android 4.3's Hidden App Permission Manager, Control Permissions for Individual Apps!* . Available: http://www.androidpolice.com/2013/07/25/app-ops-android-4-3s-hidden-app-permission-manager-control-permissions-for-individual-apps/

[31] LBE *LBE Privacy Guard*. Available: https://play.google.com/store/apps/details?id=com.lbe.security.lite&hl=en

[32] Agarwal, Y. and Hall, M. ProtectMyPrivacy: detecting and mitigating privacy leaks on iOS devices using crowdsourcing. In *Proc. of the Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 2013.

[33] Jeon, J., Micinski, K. K., Vaughan, J. A., Reddy, N., Zhu, Y., Foster, J. S. and Millstein, T. *Dr. Android and Mr. Hide: Fine-grained security policies on unmodified Android*. 2012.

[34] Nauman, M., Khan, S. and Zhang, X. Apex: extending Android permission model and enforcement with user-defined runtime constraints. In *Proc. of the ASIACCS*, 2010.

[35] Pearce, P., Felt, A. P., Nunez, G. and Wagner, D. AdDroid: privilege separation for applications and advertisers in Android. In *Proc. of the ASIACCS*, 2012.

[36] Egelman, S., Felt, A. P. and Wagner, D. Choice Architecture and Smartphone Privacy: There's a Price for That. In *Proc. of the WEIS*, 2012.

[37] Choe, E. K., Jung, J., Lee, B. and Fisher, K. Nudging People Away From Privacy-Invasive Mobile Apps Through Visual Framing. In *Proc. of the Interact*, 2013.
[38] HFEDERMAN *NTIA User Interface Mockups*. Available: http://www.applicationprivacy.org/2013/07/25/ntia-user-interface-mockups/
[39] NTIA *Privacy Multistakeholder Process: Mobile Application Transparency*. Available: http://www.ntia.doc.gov/other-publication/2013/privacy-multistakeholder-process-mobile-application-transparency
[40] Balebako, R., Shay, R. and Cranor, L. F. *Is Your Inseam a Biometric? Evaluating the Understandability of Mobile Privacy Notice Categories*. CMU-CyLab-13-011, Carnegie Mellon University, 2013.
[41] Felt, A. P., Egelman, S., Finifter, M., Akhawe, D. and Wagner, D. How to Ask for Permission. In *Proc. of the HotSec*, 2012.
[42] Jung, J., Han, S. and Wetherall, D. Short paper: enhancing mobile application permissions with runtime feedback and constraints. In *Proc. of the SPSM*, 2012.
[43] Balebako, R., Jung, J., Lu, W., Cranor, L. F. and Carolyn Nguyen. "Little Brothers Watching You:" Raising Awareness of Data Leaks on Smartphones. In *Proc. of the SOUPS*, 2013.
[44] Frank, M., Ben, D., Felt, A. P. and Song, D. Mining Permission Request Patterns from Android and Facebook Applications. In *Proc. of the Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 2012.
[45] Lin, J., Xiang, G., Hong, J. I. and Sadeh, N. Modeling people's place naming preferences in location sharing. In *Proc. of the UbiComp*, 2010.
[46] Ravichandran, R., Benisch, M., Kelley, P. G. and Sadeh, N. Capturing Social Networking Privacy Preferences. Can Default Policies Help Alleviate Tradeoffs between Expresiveness and User Burden? In *Proc. of the the Privacy Enhancing Technologies Symposium*, 2009.
[47] Cranshaw, J., Mugan, J. and Sadeh, N. User-Controllable Learning of Location Privacy Policies with Gaussian Mixture Models. In *Proc. of the AAAI*, 2011.
[48] Mugan, J., Sharma, T. and Sadeh, N. *Understandable Learning of Privacy Preferences Through Default Personas and Suggestions*. Carnegie Mellon University CMU-ISR-11-112,, 2012.
[49] Kelley, P. G., Drielsma, P. H., Sadeh, N. and Cranor, L. F. User-controllable learning of security and privacy policies. In *Proc. of the Proceedings of the 1st ACM workshop on Workshop on AISec*, 2008.
[50] Wilson, S., Cranshaw, J., Sadeh, N., Acquisti, A., Cranor, L. F., Springfield, J., Jeong, S. Y. and Balasubramanian, A. Privacy Manipulation and Acclimation in a Location Sharing Application. In *Proc. of the Ubicomp*, 2013.
[51] Benisch, M., Kelley, P., Sadeh, N. and Cranor, L. Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs. *Personal and Ubiquitous Computing*, 2010.
[52] Sadeh, N., Hong, J., Cranor, L., Fette, I., Kelley, P., Prabaker, M. and Rao, J. Understanding and Capturing People's Privacy Policies in a Mobile Social Networking Application. *The Journal of Personal and Ubiquitous Computing*, 2009.
[53] *Androguard*. Available: http://code.google.com/p/androguard/
[54] Bereau, U. S. C. *Educational Attainment*. Available: http://www.census.gov/hhes/socdemo/education/index.html
[55] Ross, J., Irani, L., Silberman, M. S., Zaldivar, A. and Tomlinson, B. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proc. of the CHI '10 Extended Abstracts*, 2010.
[56] Manning, C. D., Raghavan, P. and Schutze, H. *Hierarchical Clustering*. Cambridge University Press, City, 2008.
[57] Krause, E. F. *Taxicab Geometry*. Dover. ISBN 0-486-25202-7, 1987.
[58] Lance, G. N. and Williams, W. T. Computer programs for hierarchical polythetic classification (`similarity analyses'). *The Computer Journal*, vol. 9, pp. 60-64, 1966.
[59] Hamming, R. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, vol. 26, pp. 147-160, 1950.
[60] Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, vol. 58, pp. 236-244, 1963.
[61] Szekely, G. J. and Rizzo, M. L. Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification*, vol. 22, pp. 151-183, 2005/09/01 2005.
[62] McQuitty, L. L. similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educational and Psychological Measurement*, vol. 26, pp. 825-831, 1966.
[63] R *Hierarchical Cluster Analaysis*. Available: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html
[64] Handl, J., Knowles, J. and Kell, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, vol. 21, pp. 3201-3212, 2005.
[65] Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.
[66] Dunn, J. C. Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, vol. 4, pp. 95-104, 1974.
[67] Kumaragura, P. and Cranor, L. F. *Privacy Indexes: A Surey of Westin's Studies*. CMU-ISRI-05-138, Carneigie Mellon University, 2005.

APPENDIX A.

Template of Amazon Mechanical Turk Task

Please read the description carefully and answer the questions below. HIT will be rejected if you just click through.

[**app name**][app icon]

Developer: [developer name]

Average rating: [rating] / 5.0

Rating count: [count]

Description: [description text copied from Google Play]

[App Screenshot from Google Play #1]

[App Screenshot from Google Play #2]

[App Screenshot from Google Play #3]

You must ACCEPT the HIT before you can answer questions.

Have you used this app before? (Required)

a. Yes
b. No

What category do you think this mobile app belongs to? (Required)

a. [Candidate category #1]
b. [Candidate category #2]
c. [Candidate category #3]

Suppose you have installed [app name] on your Android device, would you expect it to access your [describing permission in plain English]? (Required)

a. Yes
b. No

Based on our analysis, [app name] accesses user's [describing permission in plain English] for [explaining purpose]. Assuming you need an app with similar function, would you feel comfortable downloading this app and using it on your phone? (Required)

a. Most comfortable
b. Somewhat comfortable
c. Somewhat uncomfortable
d. Very uncomfortable

Please provide any comments you may have below, we appreciate your input!

[text box]

It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception

Marian Harbach[1], Emanuel von Zezschwitz[2], Andreas Fichtner[2],
Alexander De Luca[2], Matthew Smith[3]
[1]Usable Security and Privacy Lab, Leibniz University Hannover, Hannover, Germany
[2]Media Informatics Group, University of Munich (LMU), Munich, Germany
[3]Department of Computer Science, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany
harbach@dcsec.uni-hannover.de, emanuel.von.zezschwitz@ifi.lmu.de,
fichtnera@cip.ifi.lmu.de, alexander.de.luca@ifi.lmu.de, smith@cs.uni-bonn.de

ABSTRACT

A lot of research is being conducted into improving the usability and security of phone-unlocking. There is however a severe lack of scientific data on users' current unlocking behavior and perceptions. We performed an online survey ($n = 260$) and a one-month field study ($n = 52$) to gain insights into real world (un)locking behavior of smartphone users. One of the main goals was to find out how much overhead unlocking and authenticating adds to the overall phone usage and in how many unlock interactions security (i.e. authentication) was perceived as necessary. We also investigated why users do or do not use a lock screen and how they cope with smartphone-related risks, such as shoulder-surfing or unwanted accesses. Among other results, we found that on average, participants spent around 2.9 % of their smartphone interaction time with authenticating (9 % in the worst case). Participants that used a secure lock screen like PIN or Android unlock patterns considered it unnecessary in 24.1 % of situations. Shoulder surfing was perceived to be a relevant risk in only 11 of 3410 sampled situations.

1. INTRODUCTION

Current mobile devices are touch-based, rich in functionality and provide high memory capacity. While early devices needed key locking mechanisms solely to prevent accidental use, current smartphones require protection mechanisms due to the potentially vast amount of private data contained on the phone. As a consequence, authentication on mobile devices has become indispensable and more secure (un)lock screens were introduced. Besides traditional alphanumeric passwords and PINs, current smartphones provide graphical as well as biometric authentication mechanisms.

Research concerning mobile authentication is also very active. One of the most cited dangers for smartphone unlocking mechanisms are shoulder surfing attacks (e.g. [3, 23, 28]). That is, direct observations with and without technical equipment (e.g. camera) aiming to capture a user's password. Based on this assumption, most proposed unlock mechanisms pay particular attention to being resistant against shoulder surfing and consequently accept reduced usability (e.g. [2, 9, 20]).

Interestingly, even though shoulder surfing is often assumed to be a relevant real-world problem, there is almost no data on the occurrence of shoulder surfing attacks in the wild or on users' perceptions of the threat. Furthermore, since lock screen mechanisms are often tested in lab environments, little is known about the users' perceptions and their behavior in real-world situations. Amongst others, important research questions are: How often and in which situations do people use secure lock screens? How often and in which context do people access sensitive data using their phone? How often is this data perceived to be in danger? And to what extent is shoulder surfing perceived to be an issue in everyday mobile device authentication?

To shed light on these questions, we conducted an online survey (n=260) and a field study (n=52), analyzing users' risk perception and behaviors when interacting with smartphone unlock mechanisms. We gathered in-depth insights into the assessment of shoulder-surfing risks and shed light on users' perceptions and daily needs when protecting their smartphone. Our approach allows us to provide a quantitative analysis of real-life unlocking behavior.

Some of our key findings are that users spend up to 9.0 % of the time they use their smartphone on dealing with unlock screens, that a secure lock screen is considered unnecessary in 24.1 % of the situations we sampled, and that shoulder surfing is only perceived to be a relevant risk in 11 of 3140 sampled situations. We also show a very diverse set of justifications for (not) having a secure lock screen, a plethora of physical measures users take to protect their phone, and that losing the smartphone-hardware is the most relevant threat to users.

We believe that the understanding gained from our studies needs to play an important role in the design of future unlocking mechanisms, since the usability/security trade-offs of current mechanisms do not match users' concerns.

2. RELATED WORK

There are two main areas of related work relevant to this paper. We will first outline the very active field of smartphone lock screen research to motivate the need for ground truth on the threats users face in their daily lives. Then, we

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

discuss existing work on the perception of security measures as well as existing data on the use of security measures on smartphones.

2.1 Unlock Screens

Authentication on mobile devices can be divided into implicit (e.g. [17]) and explicit approaches (e.g. [32]). In addition, there are mixed approaches (e.g. [7]) which add implicit security layers to an explicit authentication challenge. Implicit authentication mechanisms analyze specific time spans of behavioral cues like sensor data and usage patterns to establish a continuous authentication and hence reduce authentication workload. Examples include analyzing gait patterns [30], typing behavior [5], file system access [33], or a combination of factors [26]. Due to noticeable delays, many of them are not suited for direct lock screen mechanisms. Explicit authentication methods can be divided into biometric, token-based and knowledge-based methods [25]. The latter face the threat of shoulder surfing attacks [23].

As a consequence, the goal of finding shoulder surfing resistant solutions for knowledge-based unlock screens has become a very active research area (e.g. [3, 23, 28, 31]). Proposed concepts achieve shoulder surfing resistance either by establishing secret channels [2], by utilizing indirect input [9, 8, 20, 21], by obfuscating the input [34], or by adding additional biometric layers [7, 29].

Developing usable authentication mechanisms, which are secure against attacks such as shoulder surfing is believed to be very important. Nevertheless, to date there is no evidence that the often postulated threat of shoulder surfing attacks holds true in the users' daily lives. All of the above works were evaluated in laboratory settings and established concepts like PIN or patterns solely serve as a baseline. User perception and field performance (even of PIN and patterns), however, remain relatively unexplored. The only published work in this area focused on a quantitative performance analysis of PIN and patterns, but did not analyze real lock screen interactions [32].

Karlson et al. [18] already argued for better support of phone sharing through non-binary locking mechanisms. Prototypes of context-aware or selective authentication mechanisms for smartphones have also been proposed by Hayashi and colleagues [12, 13]. They report being able to reduce the number of authentications by up to 68 %. To date, however, there is only limited data on how these mechanisms relate to users' needs during their everyday smartphone use and which factors drive users' decisions for or against an authentication mechanism. We provide further evidence for the advantages these approaches can have, not only with respect to user workload but also to reduce the attack surface for shoulder surfers.

2.2 Security Perception and Smartphone Use

The core principle of usable security is that security is not the primary goal for regular users of computer systems [27]. Work by Beautement et al. [1] as well as Herley [14] investigated the notion of the "compliance budget" in corporate environments. According to this theory, users have a limited budget for complying with security measures and will evaluate if it is worth spending some of their budget given a particular benefit of a measure. The authors argue that users make a rational choice in rejecting the considerable number of available protection measures given a general lack of tangible benefits. However, the theories presented in these papers do not take the changing context of mobile phone use into account, where users may want to have a protection measure in one situation but not the other.

There also have been several non-academic studies that report how frequently users interact with their smartphones. For example, a study by lock screen advertising provider Locket finds that the users of their app unlock their phones 110 times a day on average[1]. In a recent market research study by Nielsen[2], researchers found that smartphone users in the UK spent almost 42 hours interacting with their smartphones in December 2013. This figure was somewhat smaller in the U.S. (34.3 hours) and Italy (37.2 hours).

3. ONLINE SURVEY

To begin to understand how users think about smartphone locking, we conducted an online survey. The aim of the survey was to get an overview of users' concerns and motivations for locking or not locking their devices. Research questions included: Why do or do not users lock their phone? Which factors play a role in their decision making about this security measure? Which kinds of attack scenarios do users consider? Are users more afraid to lose their phone in general or that someone will actually access their data? Are there any additional measures that users frequently take to protect their phones and how do these relate to having a lock screen or not?

3.1 Method

We used Amazon's Mechanical Turk (MTurk) service to distribute the survey. While MTurk does not allow us to draw representative samples of any population, the people that participate in this service have been shown to generate meaningful results in the area of usable security [19] if appropriate precautions are taken [10]. We advertised a survey about smartphone use in daily life and offered $0.70 of compensation per successfully completed task. We asked participants to only take the survey if they have been using a smartphone regularly for at least three months. They had to prove their ownership of a smartphone at the end of the survey by scanning a QR code with their device and opening the contained link in their phone's browser. The completion code was only displayed, if the HTTP user agent string matched a known mobile browser. Additionally, we included several attention check questions throughout the survey.

The survey consisted of four main parts. First, participants were asked about their smartphone use in general, including why they do or do not use a code to lock their phone and which lock screen they use. In the second part, we captured how participants value their smartphone and which risks they consider when reasoning about their phone's security. Next, we asked participants about extra measures they take to protect their phone and in which situations they take them. In the third part, participants were asked whether or not they previously had security related incidents with their smartphone. If they indicated that someone previously had unwanted access to their smartphone, we invited them to

[1] http://www.npr.org/blogs/alltechconsidered/2013/10/09/230867952/new-numbers-back-up-our-obsession-with-phones – accessed on 07.05.14

[2] http://www.nielsen.com/us/en/newswire/2014/how-smartphones-are-changing-consumers-daily-routines-around-the-globe.html – accessed on 26.02.14.

report on the most severe case, using the critical incident technique [11]. In the last part, we collected demographics and IT experience. The questionnaire can be found in Appendix A.

We used open-ended questions to ask about extra measures, the reasons why participants do (not) lock their phone, as well as critical incidents. While there were too few critical incidents reported to justify coding, we coded the reasons and extra measures using an inductive coding approach. Two of the authors independently went through the answers and created codes. To capture as many facets of participants' answers as possible, codes did not represent complete responses, but certain common aspects, such as protection goals or likely attackers. The codeplans were then discussed and merged before both authors coded all responses, assigning multiple codes to each response. Conflicting codings were again discussed and resolved before a third coder independently coded all responses again using an improved codeplan. The final round of coding yielded no more conflicts. The final codeplan can be found in Appendix B.

3.2 Participants and Results

After pretesting the survey in the lab and on MTurk, 320 workers accepted the task in November 2013. We removed 60 response sets due to incorrect completion codes (i. e. the smartphone check failed), implausible timing, or wrong answers to two or more attention check questions. The demographics are summarized in Table 1. The participants indicated high IT expertise. Almost a quarter worked in or studied IT and 39.6 % reported the highest value when asked to rate their understanding of computers and the Internet. All indicated that they use their smartphones on a daily basis with the majority using them at least once per hour. Mobile operating systems were evenly split between iOS and Android. 51.2 % of participants indicated that they have suffered from a smartphone related incident before.

Overall, 42.7 % of participants indicated that they use some form of lock screen, including PINs, passwords or unlock patterns, but not including the "slide-to-unlock" mechanism. In the remainder of the paper, this will be referred to as "code-lock". Split by operating systems, 55.2 % of iOS users were significantly more likely to have a code-lock compared to only 30.4 % of Android users (Fisher's Exact Test (FET), $p < .001$). Of the 22 Android pattern users, only 2 had made the lines between the dots invisible.

3.2.1 Locking Behavior

We asked the 111 users that use a code-lock, how frequently they think they unlock their phone on an average day. Answers ranged from 1 to 100 with a median of 20 and a mean of 24.3 times. Our field study will show that many participants significantly underestimate their phone use. Additionally, we asked these 111 users to rate their sentiments towards locking on a 5-point scale. 64.9 % were not or mostly not concerned that someone might be shoulder-surfing their code entry. 25.5 % somewhat or fully agreed that they desire an easier way of unlocking their phone, while 69.4 % somewhat or fully agreed that unlocking their phone is easy. Yet, 46.8 % also somewhat or fully agreed that unlocking their phone can be annoying. At the same time, 95.5 % somewhat or fully agreed that they like the idea that their phone is protected. These results already show a certain ambivalence towards the code-lock mechanism.

N	260	
Age	18 – 67 years	
	median 31 years	
Gender	45.4 %	female
	54.6 %	male
Occupation	50.8 %	full-time employee
	13.1 %	part-time workers
	10.0 %	self-employed
	9.2 %	student
	7.3 %	unemployed
	9.6 %	other
IT Experience	22.7 %	have worked in or studied IT
IT Expertise	39.6 %	very high self-rating
Smartphone Use	36	months (median)
Usage Frequency	79.2 %	hourly or more often
Mobile OS	49.0 %	iOS
	48.7 %	Android
	2.3 %	Other
Lock Screen	40.9 %	Slide-to-Unlock
	33.6 %	PIN
	8.5 %	Pattern
	0.8 %	Password
	16.2 %	None
Incidents	21.5 %	phone lost
	11.9 %	unwanted access
	8.5 %	stolen
	28.5 %	broken phone, lost data

Table 1: Online study participant demographics.

3.2.2 Locking Motivation

When asked why the 111 users with a code-lock chose this protection, answers centered around four topics: protection goals, protection of information, protection in specific scenarios, and protection from attackers. An overview of the 318 code instances we tagged answers with can be found in Table 2. Participants provided a very diverse set of reasons across the four main topics. However, individual participants justified their choice using only few of the available aspects (ranging from 1 to 6 codes per participant, Mdn=1.0). While many answers were unspecific ("to protect my information"), other participants provided well reasoned answers, such as increasing the time an attacker needs to access the data. It is also noteworthy that no participant mentioned protecting login credentials or logged-in accounts directly.

We asked the 149 participants without a code-based lock why they chose not to have any protection mechanism for their phone. Table 3 provides an overview of the 236 code instances we attached to the answers. In this case, answers were mostly centered around two issues, namely inconvenience and the absence of a threat. Answers again included reasonable choices, such as choosing not to have a lock screen because the contained data is not considered sensitive by the respondent, while others were less rational, such as "I don't feel like putting a password on it".

3.2.3 Smartphone Risks

To assess which risks to the content on their phones participants are most concerned about, we asked them to select the worst thing that could happen to their phone from a list of six statements (cf. Appendix A). 52.7 % stated that losing the phone itself is worst as they would have to buy a new one. This result shows that, for many users, the monetary value of the hardware is more important than the associated privacy and security risks for accounts and data. However,

Code	Count
Protection Goal	**88**
– Controlling access to phone	32
– "Safety"/"Security"	25
– "Privacy"	15
– Protection in General	6
– Increasing difficulty of unwanted access	8
– Increasing time to recover/remote-lock phone	1
– Enable data encryption	1
Protect information	**75**
– Information in general	38
– *Private* information in general	14
– Emails/Messages	9
– Photos	4
– Other app-specific content	5
– Confidential (work) information	5
Protect from specific scenario	**62**
– Lost phone	27
– Stolen phone	20
– Unattended phone	8
– Pranks/someone "messing up" phone	5
– Misplaced phone	2
Protect from attacker	**55**
– Unspecific	32
– Unwanted person	11
– Own children	11
– Roommates	1
Other	**38**
– Protect certain action	17
– Mandatory lock screen	6
– Context (work/death)	4
– Other motivation	11

Table 2: Reasons for using a code-based locking mechanism. Bold counts are sums of sub-counts.

Code	Count
Absence of threat	**118**
– don't need security	25
– nothing to hide	23
– no sensitive data	16
– keep phone physically secured	29
– use only in private environments	11
Inconvenience	**85**
– Too annoying	3
– Takes too much time	23
– Use phone too frequently	13
– Mental burden	3
Negligence/Carelessness	**8**
Dislike Locking	**7**
Other	**25**
– locking causes problems	12
– protect phone using another measure	6
– Other reason	7

Table 3: Reasons for not using a code-based locking mechanism. Bold counts are sums of sub-counts.

such risks were mentioned second-most: 20.0 % chose losing the data that is on the phone in general as the worst possible scenario, while 11.9 % chose account abuse on a lost phone and 8.8 % data abuse on a lost phone. Only 4.2 % and 1.2 % chose app abuse and data abuse respectively on an unattended phone. It has to be noted that lock screens cannot protect devices from getting lost and data loss is usually more influenced by backup strategies than authentication mechanisms. Therefore, 26.1 % of these scenarios could probably be prevented using adequate security mechanisms. The remaining 1.2 % of participants stated a combination of these six scenarios or gave another scenario. While the figures only relate to risks participants were most concerned about, these also likely influence users' behavior most.

Participants were also asked to to rate each of the six worst case smartphone risk scenarios in terms of severity and likelihood, the two classic dimensions applied to evaluate risk. We also included a third dimension, presence, that measures how frequently this risk is on a participant's mind. While the first two dimensions can capture a "value" of this risk, the third attempts to quantify how much this value influences day-to-day decision making. A risk that is considered very important by users is not only one that is particularly severe and likely but also one that is frequently present in the users' minds. In terms of presence, all six risks were on users' minds similarly infrequently: for all six risk scenarios, 65 to 82 % of participants indicated that they think of this risk infrequently or very infrequently. A Friedman's ANOVA across the six scenarios did not yield a significant difference ($\chi^2(5) = 7.74$, $p = .17$). Similarly, the likelihood of the six scenarios happening to oneself was rated as likely or very likely only by 14 to 21 % of participants. Again, these values were not significantly different ($\chi^2(5) = 1.96$, $p = .85$). There was, however, a highly significantly different rating of risks in terms of severity ($\chi^2(5) = 62.17$, $p < .001$): Post-hoc tests with Bonferroni correction revealed that losing the phone and having to replace it was considered more severe than losing data or having unwanted access to the phone. In addition, participants believed that risks to data and accounts are more severe when a phone is lost compared to when the phone is unattended.

We also asked participants to compare their individual smartphone worst case to negative situations in other contexts on a 5-point numerical scale from "not as bad" to "similar" to "worse". The situations comprised losing data on their PC, losing their wallet, losing their home or car keys, getting their e-mail account hacked or someone breaking into their home. Someone breaking into one's home was rated as somewhat worse or worse by a majority of 86.5 %. Losing the key to their home or car was rated as not as bad or similar to the worst case smartphone scenario by 60.0 % and 47.7 % respectively. Also, losing data on their PC was rated as not as bad or similar by 56.2 %. Getting their e-mail account hacked or losing their wallet ranged in between someone breaking in and the three other scenarios. This indicates that users may be ready to invest as much effort into protecting their phones as they are to protect themselves from losing the key to their home or data on their PC.

We then asked participants to rate which kinds of attackers are most likely to attempt unwanted access to their smartphones. They rated four potential attackers, known malicious and known curious as well as unknown malicious and unknown curious, on a 5-point scale from very unlikely to very likely. We found a highly significant difference between the four attackers (Friedman's ANOVA, $\chi^2(3) = 40.07$, $p < .001$) and Bonferroni-corrected post-hoc tests showed that the known curious and the unknown malicious attackers where considered more likely than the two other attackers.

For those participants who rated a known attacker as neutral, likely or very likely, we also asked whether or not they

considered eight types of known persons as a potentially curious or malicious person for their rating. The most frequently chosen types of persons are outlined in Table 4.

Curious Attackers		Malicious Attackers	
Attacker	**Freq.**	**Attacker**	**Freq.**
Close Friends	73.2 %	Other known people	68.9 %
Acquaintances	54.3 %	Co-workers	29.3 %
Parents	53.0 %	Acquaintances	25.0 %
Children	51.8 %	Friends of friends	23.2 %
Friends of Friends	46.3 %		

Table 4: Kinds of persons respondents considered as known malicious or curious attackers.

3.2.4 Extra Measures

To see how participants cope with risks to their smartphone besides inbuilt protection measures, we asked them if they sometimes apply extra measures to protect their phone. 83.5 % indicate that they keep the phone on their person or in their bag, 50.8 % leave the phone in a safe place and 33.5 % enable a lock screen or choose a harder unlock code for certain situations. Furthermore, we asked if participants with code-lock screens take some of five measures against shoulder surfing: 27.7 % indicated that they tilt their screen away while entering their unlock code when shoulder surfing is possible, 16.2 % wait a moment, 11.2 % turn around, 8.8 % cover phone, and only 7.3 % have previously changed their unlock code after a potential shoulder surfing happened. We also prompted participants to give up to three situations in which they apply those measures. As participants often not only listed a situation but also additional measures, we coded these responses for both concepts. We attached 701 instances of situation codes and 248 instances of measure codes, while each answer could receive multiple measure and situation codes. The corresponding codeplans can be found in Appendix B.3 and B.4.

In addition to the protection measures we already asked about, the coded responses revealed that in 45 instances participants mentioned to be paying extra attention to their phone. In 19 instances, other technical measures, such as turning the phone off, encrypting data, relying on remote wiping and locking functionality, removing the memory card or having a backup were quoted. With respect to situations, we found that most participants referred to public or semi-public spaces as situations where they would need extra protection. Examples include being "out" in general (59), going to events or concerts (23), while being at a gym or during workout (42), during parties or in bars (35) or at work (52). A feeling of unfamiliarity or unknown spaces were mentioned in 50 instances as were discomforting spaces, such as dark areas or dangerous neighborhoods (24). However, private spaces, such as a home, were also perceived as situations were extra measures may be necessary (16). Leaving the phone in the car (21) or uncontrolled situations where a phone is left unattended or one is less cautious (102) were frequently mentioned. In addition to unspecific unattended situations (71), participants mentioned leaving the phone to charge, while sleeping or drinking or when bags are handed over for example at the airport. Persons were also often a component of situations that were protected with extra measures (overall 61 instances): unfamiliar or untrusted persons (20), other people in general(15), kids (9), (ex-) partners (4), friends (6), and coworkers (2) were all mentioned. Finally, device sharing (5) or having sensitive and inappropriate data (4) were also quoted as situations were extra measures need to be taken.

3.2.5 Critical Incidents

The 31 participants who reported having been victim of unwanted access before, quoted the following critical incidents during which unwanted access happened: children or siblings accessing the phone for fun, snooping (ex-)partners, friends playing pranks and abusing accounts, a thief acquired the phone, friends snooping on private information, a stolen phone that was sold and then returned to the police by the buyer because the phone was not wiped, parents "checking" on their children, and having a virus on the device. We then explicitly asked about the harm that arose in this situation: ten participants stated an invasion of privacy, four got into a conflict with the other person, accounts were abused in three cases, others were offended in three cases and embarrassment was caused in one case. Seven participants reported that they were frustrated or mad and six participants indicated that they saw no harm in this incident. On the other hand, we asked participants what good came from the incident. Responses included clarified relationships and boundaries in five cases, a new phone in one case, five participants stated to have learned to pay more attention to their phone (even though they are still not locking their phone) and one started using a lock mechanism. For eight participants, nothing good came from the incident. In terms of having a code-lock or not, these critical incidents show that many of them could have been prevented by using a code-lock. However, as the previous subsections have shown, a large number of reasons let users choose not to have a code-lock.

4. LONGITUDINAL FIELD STUDY

While the survey results already provide interesting insights, they are based on self-reports at one point in time. To further evaluate the role of context to unlocking and hence generate ground truth for improvements of smartphone locking schemes, we conducted a longitudinal field study with 57 participants over four weeks. The design of the study was governed by three research questions: How frequently do people unlock their phone? What is the influence of context on perceived necessity of locking? And how frequently are users potentially subject to shoulder surfing or unwanted access to their device?

To increase data validity, we instrumented users' private phones to implement an experience sampling method and gather quantitative data like unlock frequencies and authentication times. The field study was grounded on the results of the online survey and a focus group. We conducted this focus group (n=7) to familiarize ourselves with participants' reasoning and views on our research questions. The results helped us to further reduce the question list to the most important aspects and keep the participants' additional effort as low as possible.

4.1 Method

To elicit a longitudinal picture of users' everyday behavior and perceptions, a subtle and low-effort data collection method was necessary. We decided to collect data from users of the Android OS, as it is both very common and

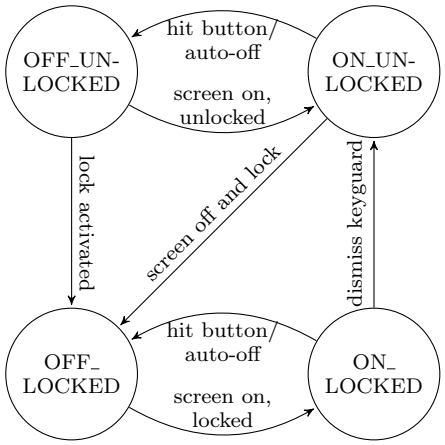


Figure 1: The states and transitions logged during data collection.

provides suitable APIs to collect the desired data. We implemented an app that would automatically log (un)locking activity on users' phones. Additionally, we displayed mini-questionnaires on random occasions to obtain a sample of users' views on their locking behavior immediately and within a given situation (cf. Section 4.1.3 below for details). The logged information was periodically backed up to our servers when the phone was connected to a WiFi network. We collected data over a period of four weeks.

Presenting questionnaires in-situ is known as the experience sampling method (ESM) and has been previously applied to investigate real-life situations [16, 6, 15]. Other longitudinal methods have been used to capture user experience on mobile phones [22], such as the Day Reconstruction Method. However, for our exploration of smartphone locking behavior and perception, we can easily use the capabilities of modern smartphones to collect the necessary data in situ and do not need to let users remember parts of their experience. Additionally, Möller et al. have previously demonstrated problems with relying on self-reporting during long-term studies [24]. We hence split our data collection efforts in two parts: Activity Logging and Mini-Questionnaires. The questionnaires only ask for immediately observable information or information from the near past and hence do not need to heavily rely on participants' memory. We present the details of our approach in the following two subsections.

4.1.1 Activity Logging

Our app monitored SCREEN_ON and SCREEN_OFF intents as well as the KeyguardManager state provided by the Android OS. This allows us to derive when a device was activated, unlocked and deactivated. Figure 1 provides a state-machine representation of the collected state information. Whenever a users presses the hardware button activating or deactivating the smartphone's screen, the state transitions between ON_* and OFF_* states. When the lock screen is dismissed, the system transitions from ON_LOCKED to ON_UNLOCKED. Finally, the transition from *_UNLOCKED to *_LOCKED occurs either immediately or after a certain delay, depending on users' configurations. We logged timestamps when entering a state. It is important to note that especially the time it takes to unlock the phone (transitioning from ON_LOCKED to ON_UNLOCKED) is a worst-case estimate, as it includes the time users spent viewing notifications or the clock on the lock screen first. Also, our app did not need any permissions to collect this data.

4.1.2 Mini-Questionnaires

As we aimed to capture participants' perceptions of threats related to their smartphone locking behavior in their daily life, we enriched the automatically logged data with participants' subjective views. We applied a method similar to what Cherubini and Oliver proposed [4]. Using two very short questionnaires, participants were asked about their surroundings and subjective perceptions. The two questionnaires were randomly displayed with a certain probability after a subset of device unlocks and contained multiple-choice questions to facilitate rapid answering. One questionnaire focussed on the unlock procedure and gathered shoulder surfing possibilities, who an attacker would be, as well as how likely and severe such an attack would be. Participants were instructed to briefly consider their environment and indicate if someone was able to see the contents of their screen in this situation. Additionally, we elicited satisfaction with the locking procedure in this situation and the sensitivity of the data to be accessed. Participants were instructed to judge sensitivity of data subjectively without giving them any further definition in order to not disrupt their own mental model. The second questionnaire focussed on the time span between the current unlock and the last use. This questionnaire elicited views on the necessity of the lock screen, if unwanted access has been possible, and how annoying the locking mechanism was in this situation. Both questionnaires asked participants to characterize the environment they are currently in as private, semi-public or public, according to the categories we obtained in the online survey as well as the pre-study focus group. The contents of both questionnaires can be found in the Appendix C.

4.1.3 Situation Sampling

To obtain a representative sample of day-to-day situations, we needed to randomly choose unlock events throughout the day after which we would display one of the two mini-questionnaires. Pre-testing showed that unlocking behavior varies widely between participants, days, and time of day. We hence dismissed the possibility to apply a fixed sampling schedule for all participants. Some participants may use their device more frequently during the day, while others may become particularly active in the evening. Additionally, we aimed to sample as many different situations as possible and therefore did not want to restrict the sampling time frame to, for instance, working hours as has been previously done in similar contexts [15]. Pre-testing also revealed that it takes about 30 to 40 seconds to complete the mini-questionnaires on the device. In order to not overwhelm participants, one of the two questionnaires would be randomly displayed with a certain probability and at most once per hour. Participants were also able to press a "Not Now" button, that would dismiss this questionnaire immediately, in order to allow quick access to the phone if necessary.

At deployment time, the probability that a questionnaire was shown for a given unlock was set to 20 % based on a one week pre-study. After one week of data collection, probabilities were adjusted to collect about 5 to 6 questionnaires per day to keep the task as unobtrusive as possible while

covering a wide range of situations. Heavy users (at least 9 unlocks per hour) were throttled to 10 % probability and medium users (between 4 and 8 unlocks per hour) to 15 %. We chose to adapt the sampling rates to put an even burden on all participants and make the study less intrusive.

4.1.4 Briefing and Debriefing

All participants were briefed about the study and the method during an initial meeting in person or by phone. The data collection procedure and the questions in both questionnaires were explained and participants had a chance to ask questions. The app was then installed on each participant's phone before participants tested both mini-questionnaires. After the data collection period, participants came in for a debriefing interview. We collected the data from participants' phones and removed all traces of the app. We also conducted a short interview, whose structure and results will be presented in Section 4.2.3.

N	52	
Age	19 – 32 years, median 23 years	
Gender	23	female
	29	male
Occupation	47	undergrad or grad students
	5	PhD student or staff
Highest degree	34	high school diploma or less
	18	Bachelor/Master degree
IT experience	25	work(ed) in or study(ed) IT
Smartphone history	34	months (mean)
Lock screen type	13	PIN
	22	Pattern
	17	Slide-to-unlock
Code lock for	22	months (mean)
Avg. PIN length	4.5	digits (range: 4-6)
Avg. Pattern length	5.2	cells (range: 4-8)

Table 5: Longitudinal field study participant demographics.

4.1.5 Participants

We recruited 57 participants at two locations in Germany, Hannover and Munich, in January 2014. At one location, 27 participants were recruited through message boards, social networks, and mailing lists, while at the other 30 students and graduates where recruited using a study participation mailing list. We advertised a four week study on Android lock screens for users that have had a smartphone with Android 2.3 or higher for at least 3 months. A 10 Euro base-salary plus 14 Euro-cent per completed mini-questionnaire were promised as compensation. Participants earned 30.79 Euros on average.

While all 57 participants completed the data collection part of the study, we removed one participant who did not show up for debriefing, three participants who repeatedly modified the time on their phone during data collection, and one participant where data collection failed for several days, as our app did not restart after rebooting this user's device. The remaining 52 participants' demographics are summarized in Table 5.

While the participants mainly comprise students of which about half also have some IT experience, we believe that this is a population worth studying as they are often very active experiencing a wide range of situations but also have phases where they sit in front of a desk for extended periods of time. As we aim to explore how different environments influence locking behavior and risk perception, our sample offers a good chance to collect a wide range of usage contexts. However, it still has to be noted that the results cannot be generalized to any particular population.

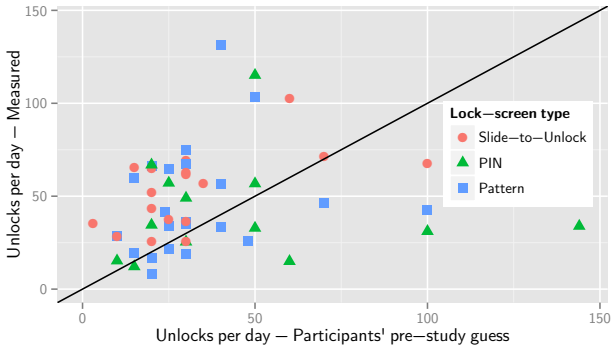


Figure 2: A comparison of participants' pre-study guess of unlocks per day versus the actually measured values.

4.2 Results

Participants contributed 29.5 days of data on average. To equalize the time we analyze per user, we pruned each participants' dataset to 27 complete days from midnight to midnight by removing the first hours and the remaining days. Due to our method, each user contributed a different amount of data. In order to not over-represent users that use their phone more frequently, we first aggregate data per user and then average across users' aggregates where appropriate.

4.2.1 Logged Data

Within the 27 days, we observed an average of 2242.3 activations (switching the screen of the device on) per participant (sd = 1160.2, median=2260), ranging from 651 to 5419. Correspondingly, 1286.0 unlocks (dismissing the lock screen after activating the phone) were logged on average per participant (sd = 711.8, median=1127), ranging from 215 to 3545.

Per day, participants activated their phone 83.3 times (sd = 43.0, median=83.8) and unlocked 47.8 times (sd = 26.4, median=42.1) on average. This translates to an average of 5.2 activations and 3.0 unlocks per hour, assuming that a user is awake for 16 hours per day. Participants unanimously attributed the discrepancy between activations and unlocks to activating the screen of their phone to see the current time and to check for notifications. Overall, usage was largely similar during daytime hours, ramping up in the morning and down in the evening after 9 pm (also cf. Figure 8 in the Appendix).

During recruitment, we asked participants how frequently they think they unlock their phone per day. Figure 2 compares these guesses with the measured frequency. We find that most users severely underestimated their use. However, participants who use their phone less frequently appeared to give better estimates.

Figure 3 shows that the distribution of unlocks per hour across users is bimodal. We hence group users into heavy

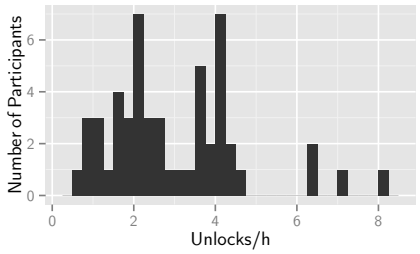


Figure 3: Histogram of users' mean combined activation and unlock times.

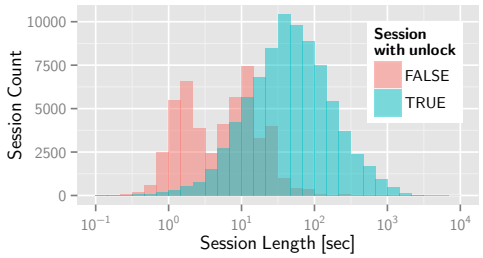


Figure 4: Histogram of session lengths on a log scale.

and "regular" users, where heavy users unlock their phone more than 3 times per hour. Please note that significance testing results based on this grouping are only of exploratory nature, as groups were formed post-hoc.

Activating and unlocking the phone took 2.67 seconds without a code lock (sd = 8.46s, median=1.26s), 3.0 seconds using a lock pattern (sd = 13.3 sec, median=1.69s), and 4.7 seconds using a numeric PIN (sd = 20.72s, median=2.85s) across all unlocks. Averaging unlock times per user, we ran a user-type by lock-type between-subjects ANOVA and found a highly significant main effect for lock-type ($F(2, 46) = 11.37$, $p < .001$) as well as a significant main effect for user-type ($F(1, 46) = 6.39$, $p = .002$). Heavy users completed their unlocks more quickly on average (2.9 vs. 3.8 seconds). Holm-corrected pairwise testing also showed that PIN (4.9 seconds on average) was significantly slower than the two other mechanisms (Slide-to-Unlock 2.6 and Pattern 3.2 seconds, $p < .001$, Cohen's d = 1.58 and 1.27 respectively). During the 27 days of the experiment, participants spent an average of 1.17 hours each (sd = .87, ranging from .2 to 5.1 hours) just unlocking their device. There also was a significant correlation between unlocking time and unlocking frequency (Spearman's $\rho = -.30$, $p = .034$).

An average session (from SCREEN_ON to SCREEN_OFF) lasted 70.3 seconds (sd = 241.5s). However, sessions where participants actually saw the home screen lasted for 104.1 seconds (sd = 193.9s, median=45.6s) on average, including the time it took to dismiss the lock screen. The remaining sessions (those when the device was not unlocked) lasted only 12.4 seconds (sd = 297.6s, median=5.2s) on average. Figure 4 gives an overview of session lengths, grouped by whether or not the session entered the home screen. It can be clearly seen that sessions last longer once the lock screen was dismissed. Also, the distribution of session lengths on a locked device is bimodal. We hypothesize that the maximum at about one second is for checking the time, while the maximum at about 10 seconds session lengths represents cases where users check notifications. Averaging per user, we did not find a significant correlation between unlock frequency and average session time.

Overall, users spent 43.0 hours on average (sd = 22.1h, median=41.2h) using their smartphone within the 27 days of our experiment, of which an average of 2.9 hours were spent on a locked device (i. e. checking time or notifications on the lock screen). 2.9 % of the overall time was related to unlocking the phone on average, ranging from .6 to 9 %.

4.2.2 *Questionnaire Data*

We collected 3410 completed unlock risk questionnaires (65.6 per user on average, range 15-110) and 3172 completed data risk questionnaires (61.0 per user on average, range 15-105). The sampled situations included a wide range of times of day and even collected samples when participants used their phone at night (cf. Figure 8 in the Appendix). Filling the questionnaires took 23.7 (sd = 35.9) and 21.3 (sd = 22.6) seconds on average for each type respectively. In the following, we present results from questionnaire parts individually.

Environments.

In both questionnaires, participants reported the environments in which they were in the moment they unlocked the phone or in which they have been since they last used the phone. Averaging environment proportions per user, these environments were mostly private (62.4 %), semi-public in 19.5 % of cases and public in 18.2 %. In line with previous findings [12], this indicates that most smartphone use takes place at home or in similarly private spaces.

Perception of Lock Screen.

In the first mini questionnaire, we asked participants how annoying the unlock (which they just completed prior to filling out the questionnaire) was. Participants reported different proportions of annoying unlocks (either "annoying" or "very annoying"). Figure 5 shows the relationship between the proportion of annoying unlocks, the number of completed questionnaires (corresponding to how heavily users use their smartphone), and the type of lock screen they use. A large amount of participants was very happy with their lock screen, as they reported no or almost no annoying unlocks across their questionnaires. Only 12 of 52 participants indicated being annoyed by their lock screen in more than 50 % of their mini-questionnaires. There also is no clear trend of users with a particular lock type being more annoyed. However, we note that only three users with Slide-to-Unlock reported annoying unlocks in more than a quarter of their questionnaires.

Additionally, in the other mini questionnaire, we asked if users with a code lock would have rather not have had a code lock in this situation and vice versa. High ratings on the 5-point numeric scale of this question indicate dissatisfaction with having a code lock or not. Figure 6 and Table 6 give an overview of the answers provided. In the figure, the y-axes additionally show how many questionnaires each user completed, approximating how frequently the phone is used.

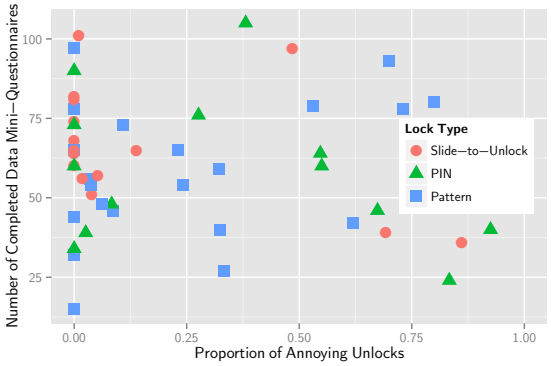


Figure 5: Proportion of annoying unlocks per user versus how many questionnaires were completed.

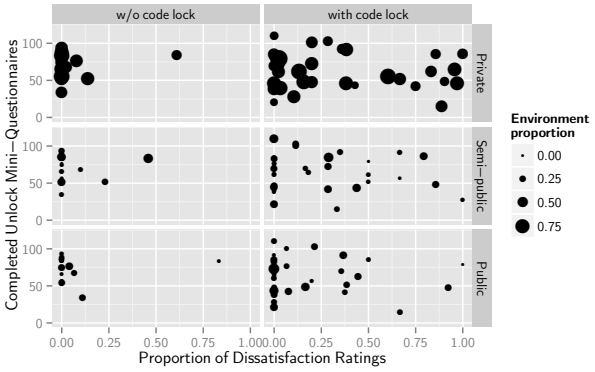


Figure 6: Proportion of dissatisfied ratings per user versus how many questionnaires they completed, grouped by whether they had a code lock screen and in which environment the rating was given.

Participants' answers are grouped by the environment they were provided in and whether or not this participant had a code-lock. The size of each point in the graph indicates how frequently this user reported being in this environment.

The data shows that participants without a code lock were generally more satisfied with their status quo. Only few of them indicated dissatisfaction in more than a quarter of their responses across all environments. Participants with code locks showed more variability and more participants indicated dissatisfaction in more than a quarter of their responses. Especially users that are frequently in private environments were very dissatisfied with their code locks. It is also noteworthy that fewer code-lock participants indicated strong dissatisfaction in public environments compared to semi-public or private situations.

A possible interpretation is that being annoyed by a lock mechanism overlays risk perception to some extent as there is only a limited trend towards more satisfaction with lock screens in potentially more dangerous public situations.

Data Sensitivity.

We asked each participant for subjective ratings on how sensitive the data that is going to be accessed in this session is. In 684 (20.1 %) of 3410 completed mini questionnaires, users indicated that they did not know what kind

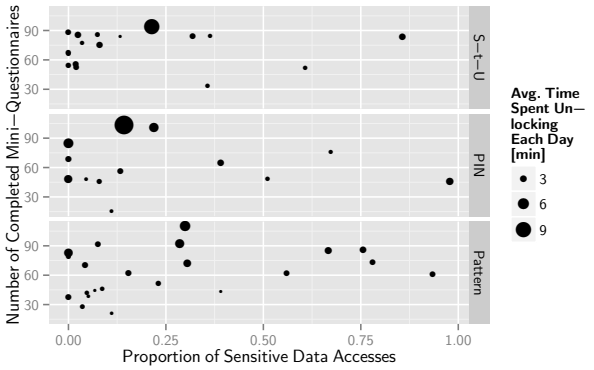


Figure 7: Proportion of sensitive data accesses per user, grouped by lock mechanism. The y-axis also shows completed questionnaires.

of data they were going to access. Aggregating proportions of unknown accesses per participant, the mean proportion amounts to 19.6 % ($sd = 25.0$ %) and participants' individual values range from 0 % to 88.2 %.

In 25.3 % (691) of the 2726 remaining reported situations, participants indicated accesses to sensitive data. For each user, this means that during the experiment only 10.6 hours ($sd = 15.0$) of the 43 hours each participant spent using their device contained accesses to sensitive data on average. All but ten users indicated that they access less sensitive data in more than half of the sampled sessions. Figure 7 visualizes the proportions of sensitivity ratings across the sampled situations per participant. It is also visible that one user spent a lot of time unlocking the phone each day even though the data that should be accessed was not sensitive in most cases. Notably, the ten participants that were accessing most sensitive data use their phone more frequently (i. e. filled more questionnaires).

Shoulder Surfing.

Table 7 gives an overview of shoulder surfing possibilities perceived by our participants. Across the 3410 unlock risk mini-questionnaires we collected, shoulder surfing was not perceived to be possible in a majority of 83.0 % of cases. When it was possible, mostly known persons were observers, except in public environments. In more than half of the situations where shoulder surfing would have been possible, participants thought it to be unlikely or very unlikely that this did actually happen. Had it happened, the threat from the potential attacker would have been low or very low in most of the possible shoulder surfing situations, especially in private environments. Overall, we found only 11 of the 3410 (.3 %) reported situations were it was likely that a shoulder surfer was looking at the screen and it would have been severe or very severe if that had actually taken place. Seven of these occurred in public situations.

We also asked those participants with a code lock whether or not they protected their code entry during the last unlock, by for example tilting their screen away from onlookers or waiting to unlock the phone. Only 18 participants reported 52 instances in which they actively protected the code input from a shoulder surfing threat within 1869 sampled situations where a code was entered (2.8 %).

Environment	# Situations	Mean Proportion of Dissatisfaction Ratings w/o code lock	with code lock	overall
private	2115 (62.0 %)	5.0 % ($sd = 14.9$ %)	32.7 % ($sd = 36.0$ %)	23.6 % ($sd = 33.2$ %)
semi-public	690 (20.2 %)	4.6 % ($sd = 12.2$ %)	23.0 % ($sd = 29.3$ %)	17.0 % ($sd = 26.3$ %)
public	605 (17.7 %)	6.2 % ($sd = 20.1$ %)	16.6 % ($sd = 26.9$ %)	13.2 % ($sd = 25.2$ %)
Overall	3410	5.3 % ($sd = 15.8$ %)	24.1 % ($sd = 31.4$ %)	17.9 % ($sd = 28.6$ %)

Table 6: Participants' dissatisfaction with their locking mechanisms by environment.

Environment	# Situations	Known Person	Unknown Person	Nobody	Unlikely	Low Severity
private	2115 (62.0 %)	8.6 % (181)	0.0 % (1)	91.4 % (1933)	56.6 % (103)	92.9 % (169)
semi-public	690 (20.2 %)	22.2 % (153)	4.6 % (32)	73.2 % (505)	65.4 % (121)	84.9 % (157)
public	605 (17.7 %)	10.4 % (63)	24.5 % (148)	65.1 % (394)	56.0 % (118)	68.3 % (144)
Overall	3410	11.6 % (397)	5.3 % (181)	83.0 % (2832)	59.2 % (342)	81.3 % (470)

Table 7: Shoulder surfing possibilities across potential "attackers" and environments. The last two columns give percentages with respect to possible shoulder surfing attempts (i. e. by known or unknown persons).

Unwanted Access.

In the data risk mini-questionnaire, participants were asked to report situations in which unwanted access to their smartphone was possible. Eleven participants did not report any of these situations and the remaining 42 participants reported a total of 245 occasions out of 3172 possibilities (7.7 %) and between one and twenty occasions each. Table 8 provides an overview of unwanted access occasions, who an attacker would have been, how many of these occasions were rated as unlikely and for how many the consequences participants saw were rated as benign. Unwanted accesses were infrequently possible, mostly by known persons except in public situations and rated as mostly unlikely and benign.

4.2.3 Debriefing Interview

During the debriefing sessions, we asked if participating in the study or seeing the questionnaires influenced participants' smartphone use. One participant reported to have increased the time interval after which the lock screen is shown again from 30 to 90 seconds, another participant stated that he sometimes did not turn his screen off immediately. Three participants stated that they may have used the device a little less frequently at the beginning of the study. Ten participants said that being part of the study made them pay more attention to why and how often they use their phone. While it made them realize their usage, they reported not to have altered their behavior. One participant said that he may remove his code-lock after the study, as participating made him realize how much effort unlocking with a PIN takes.

Participants were also asked to rate how annoying they found answering the mini-questionnaires to be. Only 5 participants selected 4 on a numeric scale from not annoying at all (1) to very annoying (5). 43 participants chose 2 or 3 and an additional 4 chose not annoying at all. On the contrary, many users reported that they found participating in the study very interesting for themselves, as it helped them assess their own behavior better. We also presented participants with a summary of the data they had shared with us, including frequencies of logged events, general usage statistics as well as overviews of mini-questionnaire answers. Most participants found these figures to be interesting and sometimes alarming, as they would not have expected to activate or unlock their phone as frequently. We also gave participants a numeric scale asking how well the collected data represents their actual behavior (logged data) and perception (questionnaire answers) from "not at all" (1) to "very much" (5). Participants felt that the data was valid: only one participant chose 3, 31 chose 4, and 20 chose 5.

To see how well the sampled situations covered participants' daily lives, we asked them if there were additional situations in which unwanted access was possible and if so of which nature those were. Several participants said that there probably were more of these situations, but they were mostly the same as the ones they reported in the sampled situations. Similarly, we asked participants if the proportion of situations where shoulder surfing was possible matched their own perception. Participants agreed that the numbers we collected and the proportion of shoulder surfing situations match their perception beyond the situations were questionnaires were shown. However, several participants mentioned that there were brief situations mostly in public environments where shoulder surfing would have been possible but no questionnaire was shown.

As in the online survey, we asked participants about previous critical incidents with their smartphone. Four participants had lost their smartphone before and two had unwanted access. In all cases, a lock screen was helpful to prevent more damage or was activated after the incident.

We also asked participants why they chose to have a lock mechanism with a code and coded results using the codes from the online survey. The 37 participants' answers contributed 115 code instances summarized in Table 9. The results are similar to the online survey with the exception that several participants also gave restricting statements, noting that they do not believe that lock screens offer perfect security (7), that they do not really need security (6), or that others know their code anyway (3).

Again, participants without code locks also justified their choice and 15 participants contributed 44 code instances. Table 10 provides an overview of the reasons. The most frequently cited reasons for not using a lock, as in the online survey, are inconvenience and not seeing a threat.

Finally, we asked how sensitive participants consider the data on their smartphones and whether or not they share their code with other people. 22 participants (42.3 %) chose sensitive or very sensitive on a 5-point scale, while 23.5 % of users without a code-lock and 48.6 % of users with such a

Environment	# Situations	Known Person	Unknown Person	Unlikely	Benign Cons.
private	131 (53.5 %)	97.7 % (128)	2.3 % (3)	92.4 % (121)	86.3 % (113)
semi-public	75 (30.6 %)	70.7 % (53)	29.3 % (22)	93.4 % (70)	64.0 % (48)
public	39 (15.9 %)	23.1 % (9)	76.9 % (30)	79.5 % (31)	18.2 % (11)
Overall	245 (7.7 %)	77.6 % (190)	22.4 % (55)	90.6 % (222)	70.2 % (172)

Table 8: Unwanted access occasions by environments and potential attackers. The last two columns give percentages of likelihood and severity of consequences with respect to reported unwanted access occasions.

Code	Count
Specific protection goal	13
Unspecific protection goal ("Security")	12
Specific attacker	13
Unspecific attacker	10
Protect from specific scenarios (e.g. lost, stolen)	20
Protecting specific information	5
Protecting unspecific information	8
Protect from accidental input	4
Custom certificate	5

Table 9: Reasons for using a code-based locking mechanism of field study participants.

Code	Count
Inconvenience	17
Absence of threat	16
Locking causes problems	6
Protect phone using another measures	4
Not secure anyway	2

Table 10: Reasons for not using a code-based locking mechanism of field study participants.

mechanism considered the data on their smartphones to be sensitive. However, this difference is only almost statistically significant (FET, $p = .076$). Unlock codes were shared with at least one person by 28 of 35 participants with a code-lock. Six participants indicated that at least 5 other people know their code. This also indicates that code-based locking mechanisms can be problematic in device sharing situations, as already noted by Karlson et al. [18].

5. DISCUSSION

In the two previous sections, we presented results from two studies, which we summarize and discuss grouped by the most important observations in the following sections.

5.1 High Number of Unlocks

36 of 52 participants underestimated the number of smartphone unlocks by 141 % on average. This indicates that unlocking is a subliminal action in many cases and unlock effort is kept low enough most of the time. However, even if a single unlock took only between 2.67 seconds (slide to unlock) and 4.7 seconds (PIN), the huge number of daily unlocks leads to a high impact of every additional second. Just over the course of our experiment, participants on average already spent about one hour unlocking their devices using traditional unlock screens. Taking into account that alternative authentication mechanisms often incur higher input times for increased security, this can easily add several hours of additional unlock time per month. This is especially critical when considering that average usage times per activation are relatively short and shows that authentication speed of feasible systems must be about as fast as PIN and patterns.

Since out data indicates that unlocks are perceived as unnecessary in private environments and sensitive data is seldom accessed, we suggest that more effort should be put into researching how to decrease the number of unlocks by deploying usable context- and content-dependent locking mechanisms. The work of Hayashi et al. [13, 12] are a first step in this direction.

5.2 Reasons for (Non-)Use of Authentication are Highly Diverse

The results of both studies suggest that reasons for using or not using protection mechanisms to access smartphones are highly diverse. Often, they are not based on objective reasons and were not valid from a technical perspective. In turn, a considerable number of participants provided reasonable justifications. Furthermore, others argue that code locking mechanisms are not perfectly secure anyway and even have drawbacks should the device be lost.[3] Participants without a code-lock in the field study were also very satisfied with their choice and indicated very few situations where they would have rather had a lock screen. In turn, dissatisfaction with a code-based lock was not as pronounced in public situations, as participants valued protection slightly more in that case. In terms of attackers, survey participants were most afraid of unknown malicious as well as known curious attackers. This is mirrored in the field study results, where known persons had the most shoulder surfing and unwanted access possibilities in private environments while unknown persons dominated in public situations.

5.3 Protection is More Than Authentication

Throughout the analysis, it became apparent that most participants who did not use authentication to protect their phone did not consider themselves to be unprotected. We were able to identify a fair number of approaches that participants applied to protect their devices in the online study. These users felt secure despite the absence of authentication. For instance, participants reported to never leave their devices unattended when in public settings and to keep them close at all times (e.g. in their pockets or bags). This is also mirrored in the low number of high impact unwanted access possibilities during the field study.

This is even more interesting when analyzing the risks related to smartphone use. Only 26 % of the perceived worst-case risks in the study could actually be avoided by authentication. These included risks like theft or loss of the device itself. In many cases, participants rated the monetary value of their devices higher than the possibility of losing their

[3]The finder is not able to access the address book to find the owner.

data or someone gaining access to the data. Similarly, absence of threat was a very frequently mentioned reason for not having a lock screen in the online and the field study.

5.4 Sensitive Data is Seldom Accessed

As mentioned before, when filling out the questionnaire, participants were asked whether the accessed data is sensitive for them. This was the case in 25.3 % of all sampled unlocks. This means that nearly 75 % of interactions with the smartphones were with non-sensitive data. Taking into account the overhead created by the authentication process, there is high potential for lowering the burden for the users. That is, the results indicate that binary authentication as we are using it today (i.e. all or nothing access to a device) should be seriously re-assessed. For instance, instead of protecting the mobile operating system in its entirety, protection might be used on a data level. We can see a current trend in the mobile phone industry, granting access to non-sensitive functionality like flashlight and camera (not photos) without the need for protection. Our results suggest that this does not go far enough and more aspects of the phone could be used without the need for authentication. Hayashi et al. [13, 12] already proposed potential solutions for this problem.

5.5 Shoulder Surfing Risks Perception

The results of the field study indicate that the perceived shoulder surfing risks are rather low. Our participants believed shoulder surfing would have been possible in 17 % of reported cases. However, it was considered a high risk in only 11 out of 3410 occurrences. Additionally, participants protected themselves against such attacks using physical measures only in 2.8 % of sampled situations. Overall, we can state that the participants were aware of possibly risky situations but that this did not influence their general opinion about protecting against this threat. While shoulder surfing can take place in any environment, unknown attackers are mostly present in public environments, which were however frequented least by our participants.

Shoulder surfing in private environments was mostly considered possible by people known to the user. This was, however, often not considered a threat or those people knew the lock codes anyway. Yet, this does not mean that shoulder surfing is not a risk worth addressing by improved technology. Just because users do not perceive a threat as serious does not mean that it is not. It does however mean that the additional effort a user is willing to invest to protect from it needs to be carefully assessed. Based on our results we also recommend that the shoulder surfing attack risk can be minimized by reducing the number of "unnecessary" code entries. Since shoulder surfing resistant authentication mechanisms often incur reduced performance, the user should be able to decide in which situation protection is actually necessary.

6. ETHICAL CONSIDERATIONS

While there is no reviewing board at the involved institutions for this type of user studies, all studies have to comply with federal law and privacy regulations. We conducted both studies in compliance with these strict rules. For example, identifying information had to be removed from the data before analysis and participants can only be identified in cases for which they gave explicit consent (for example to receive their compensation).

7. LIMITATIONS

The online study as well as the field study both have limitations. The online survey relied on self-reporting and can hence only shed limited light on real behavior. We therefore focussed this investigation on respondents' perceptions, attitudes and common practices. The field study also logged behavioral data, but uses a different sample of participants as well as a limited set of sampled situations. While a considerable number of situations was sampled across 27 days, participants also indicated that some rare occasions and situations that did not last very long have been missed. Furthermore, extreme situations caused participants to dismiss the questionnaire, as they needed to access information quickly. Showing the questionnaires also heightened participants awareness of risk and their own behavior. This may have influenced participants' responses.

Similarly, we were only able to extract certain events from the Android OS. The reported times for the duration of the unlock therefore also include occasions where participants first read their notifications and only unlocked afterwards. The reported times should therefore be treated as upper limits. However, as this behavior is likely similar across the lock mechanisms, the respective values should still be comparable.

Finally, the field study also included self-reported and subjective views. Participants may have categorized similar situations as, for example, public or semi-public environments, depending on their perception. Also, the same data may be perceived as more or less sensitive by individual participants and attack opportunities may have been missed. However, we believe it is the participants' views that count more than absolute numbers, as they are more likely to adopt improved security measures if they see a relevant threat by themselves.

8. CONCLUSION AND FUTURE WORK

We were able to provide in depth insights into users' interactions with smartphone locking mechanisms. The online survey gave a broad overview of participants' reasons for (not) using lock screens, how they protect their phones, and which critical incidents have previously happened to them. In addition, the longitudinal field study captured one month of unlocking activity and sampled 6582 situations in situ, providing reliable ground truth for further explorations.

We found that there is a massive number of unlocks that the participants themselves severely underestimated. Participants also showed very diverse reasons for locking or not locking their phone. The insights gathered from our studies can help future efforts to improve the adoption of smartphone protection mechanisms. We also demonstrated that users apply many physical measures to protect their phone, which often makes additional IT measures superfluous in the their opinion. Sensitive data was found to be seldom accessed which provides an opportunity to reduce the attack surface of shoulder surfing.

We believe that in future work, these results can be used to improve the design of unlock mechanisms for mobile devices in general and their adoption in particular. Additionally, it would be interesting to extend our study to include users with more diverse demographics to assess their needs and allow for a tailoring of mechanisms to specific audiences.

9. REFERENCES

[1] A. Beautement, M. A. Sasse, and M. Wonham. The Compliance Budget. In *Proc. New Security Paradigms Workshop (NSPW)*, 2008.

[2] A. Bianchi, I. Oakley, V. Kostakos, and D. S. Kwon. The Phone Lock: Audio and Haptic Shoulder-surfing Resistant PIN Entry Methods for Mobile Devices. In *Proc. TEI*, pages 197–200, 2011.

[3] R. Biddle, S. Chiasson, and P. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Comput. Surv.*, 44(4):19:1–19:41, Sept. 2012.

[4] M. Cherubini and N. Oliver. A Refined Experience Sampling Method to Capture Mobile User Experience. In *International Workshop of Mobile User Experience Research - Proc. CHI EA*, 2009.

[5] N. Clarke and S. Furnell. Authenticating mobile phone users using keystroke analysis. *International Journal of Information Security*, 6(1):1–14, 2007.

[6] S. Consolvo, B. Harrison, I. Smith, M. Y. Chen, K. Everitt, J. Froehlich, and J. A. Landay. Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies. *International Journal of Human-Computer Interaction*, 12(1-2):103–118, 2007.

[7] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. Touch Me Once and I Know It's You!: Implicit Authentication Based on Touch Screen Patterns. In *Proc. CHI*, 2012.

[8] A. De Luca, M. Harbach, E. von Zezschwitz, M.-E. Maurer, B. Slawik, H. Hussmann, and M. Smith. Now You See Me, Now You Don't – Protecting Smartphone Authentication from Shoulder Surfers. In *Proc. CHI*, 2014.

[9] A. De Luca, E. von Zezschwitz, N. D. H. Nguyen, M.-E. Maurer, E. Rubegni, M. P. Scipioni, and M. Langheinrich. Back-of-device Authentication on Smartphones. In *Proc. CHI*, 2013.

[10] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are Your Participants Gaming the System? Screening Mechanical Turk Workers. In *Proc. CHI*, 2010.

[11] J. C. Flanagan. The Critical Incident Technique. *Psychological Bulletin*, 51(4):327, 1954.

[12] E. Hayashi, S. Das, S. Amini, J. Hong, and I. Oakley. CASA: Context-Aware Scalable Authentication. In *Proc. SOUPS*, 2013.

[13] E. Hayashi, O. Riva, K. Strauss, A. J. B. Brush, and S. Schechter. Goldilocks and the Two Mobile Devices: Going Beyond All-Or-Nothing Access to a Device's Applications. In *Proc. SOUPS*, 2012.

[14] C. Herley. So Long, And No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proc. New Security Paradigms Workshop (NSPW)*, 2009.

[15] R. M. Hogarth, M. Portell, and A. Cuxart. What Risks Do People Perceive in Everyday Life? A Perspective Gained from the Experience Sampling Method (ESM). *Risk Analysis*, 27(6):1427–1439, 2007.

[16] S. S. Intille, J. Rondoni, C. Kukla, I. Ancona, and L. Bao. A Context-Aware Experience Sampling Tool. In *Proc. CHI-EA*, 2003.

[17] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit Authentication for Mobile Devices. In *Proc. USENIX HotSec*, 2009.

[18] A. K. Karlson, A. J. B. Brush, and S. Schechter. Can I Borrow Your Phone?: Understanding Concerns When Sharing Mobile Phones. In *Proc. CHI*, 2009.

[19] P. G. Kelley. Conducting Usable Privacy & Security Studies with Amazon's Mechanical Turk . In *Proc. SOUPS*, 2010.

[20] R. A. Khot, P. Kumaraguru, and K. Srinathan. WYSWYE: Shoulder Surfing Defense for Recognition Based Graphical Passwords. In *Proc. OzCHI*, 2012.

[21] S.-H. Kim, J.-W. Kim, S.-Y. Kim, and H.-G. Cho. A new Shoulder-Surfing Resistant Password for Mobile Environments. In *Proc. ICUIMC*, 2011.

[22] S. Kujala and T. Miron-Shatz. Emotions, Experiences and Usability in Real-life Mobile Phone Use. In *Proc. CHI*, 2013.

[23] J. Maguire and K. Renaud. You Only Live Twice or "The Years We Wasted Caring About Shoulder-Surfing". In *Proc. Conference on People and Computers*. British Computer Society, 2012.

[24] A. Möller, M. Kranz, B. Schmid, L. Roalter, and S. Diewald. Investigating Self-Reporting Behavior in Long-Term Studies. In *Proc. CHI*, 2013.

[25] L. O'Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12):2021–2040, Dec 2003.

[26] O. Riva, C. Qin, K. Strauss, and D. Lymberopoulos. Progressive Authentication: Deciding When to Authenticate on Mobile Phones. In *Proc. USENIX Security*, 2012.

[27] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the 'Weakest Link' – A Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, 19(3):122–131, 2001.

[28] F. Schaub, R. Deyhle, and M. Weber. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *Proc. MUM*, 2012.

[29] M. Shahzad, A. X. Liu, and A. Samuel. Secure Unlocking of Mobile Touch Screen Devices by Simple Gestures: You Can See It but You Can Not Do It. In *Proc. MobiCom*, pages 39–50, 2013.

[30] M. Tamviruzzaman, S. I. Ahamed, C. S. Hasan, and C. O'brien. ePet: When Cellular Phone Learns to Recognize Its Owner. In *Proc. SafeConfig Workshop*, pages 13–18, 2009.

[31] F. Tari, A. A. Ozok, and S. H. Holden. A Comparison of Perceived and Real Shoulder-surfing Risks Between Alphanumeric and Graphical Passwords. In *Proc. SOUPS*, pages 56–66, 2006.

[32] E. von Zezschwitz, P. Dunphy, and A. De Luca. Patterns in the Wild: A Field Study of the Usability of Pattern and Pin-based Authentication on Mobile Devices. In *Proc. MobileHCI*, pages 261–270, 2013.

[33] S. Yazji, X. Chen, R. P. Dick, and P. Scheuermann. Implicit User Re-authentication for Mobile Devices. In *Ubiquitous Intelligence and Computing*, Lecture Notes in Computer Science, 2009.

[34] N. H. Zakaria, D. Griffiths, S. Brostoff, and J. Yan. Shoulder Surfing Defence for Recall-based Graphical Passwords. In *Proc. SOUPS*, 2011.

APPENDIX

A. ONLINE-SURVEY QUESTIONNAIRE

Smartphone Risk Attitudes.

- **IF CODE LOCK:** Please estimate how many times you approximately unlock your phone on an average day. – *Numeric answer*
- **IF CODE LOCK:** Please briefly state why you are using a lock screen on your device. – *Open-ended answer*
- **ELSE:** Please briefly state why you chose not to use a PIN, password, or pattern lock screen on your device. – *Open-ended answer*
- **IF CODE LOCK:** Please rate the following statements concerning your lock screen. – *5-point numeric scale anchored at don't agree and fully agree.*
  - Unlocking my phone is annoying sometimes.
  - I like the idea that my phone is protected from unauthorized access.
  - It is difficult to unlock my phone.
  - I wish there was an easier way of unlocking my phone.
  - Unlocking my phone is easy.
  - I am concerned that someone might be observing my unlocking password/pattern/PIN in order to access my phone at a later time.
- What's the worst thing that could happen to your smartphone?
  - Losing the phone itself, because I would have to buy a new one.
  - Losing the data that is on my phone (e.g. photos, contacts).
  - Someone being able to access my data when I lose my phone.
  - Someone being able to abuse my accounts and apps when I lose my phone.
  - Someone being able to access my data when my phone is unattended.
  - Someone being able to abuse my accounts and apps when my phone is unattended.
  - Other: *text field*
- Please rate how the following events compare to the worst thing that could happen to your smartphone (Your answer was: <previous answer>). – *5-point numeric scale anchored at worse, similar and not as bad.*
  - Losing data on my computer
  - Losing my wallet
  - Losing the key to my home
  - Losing the key to my car
  - Getting my email account hacked
  - Someone breaking into my home
- Please rate how serious you find the following incidents. – *5-point numeric scale anchored at not serious and very serious.*
  - *same items as "What's the worst thing..."*
- How likely do you believe it is that each of the following things occurs to you personally? – *5-point numeric scale anchored at very unlikely and very likely.*
  - *same items as "What's the worst thing..."*
- How frequently do you think about each of the following things? – *5-point numeric scale anchored at very infrequently and very frequently.*
  - *same items as "What's the worst thing..."*
- How likely do you consider the following groups of people to be attempting to access your smartphone? – *5-point numeric scale anchored at very unlikely and very likely.*
  - Unknown malicious person
  - Unknown curious person
  - Known malicious person
  - Known curious person
- **IF known person considered likely:** Which of the following groups of known people did you just consider as potentially interested in accessing your phone without your permission? – *Choice from: Potentially curious person, potentially malicious person, I did not consider this group of people.*
  - Acquaintances
  - Close friends
  - Friends of friends
  - Parents
  - Children
  - Other relatives
  - Co-workers and colleagues
  - Other people

Extra Measures.

- Do you sometimes take additional measures to protect your smartphone in particular situations? – *Choose all that apply.*
  - I leave my phone in a safe place before going somewhere.
  - I conceal my smartphone in my clothes or in a bag.
  - I enable a lock screen for this situation or choose a harder PIN/password/pattern.
  - Other: *text field*
- **IF MEASURES TAKEN:** Please list up to three situations in which you sometimes take additional measures to protect your smartphone. – *Open ended answer in three text fields.*
- **IF CODE LOCK:** If you think someone is able to see the screen of your phone, do you sometimes take additional measures to protect your smartphone? – *Choose all that apply.*
  - I cover my smartphone while entering my PIN or pattern.
  - I wait a moment before entering my PIN or pattern.
  - I turn around before entering my PIN or pattern.
  - I tilt my screen away before entering my PIN or pattern.
  - I change my PIN/password/pattern after someone could have seen my screen.
  - Other: *textfield*

Critical Incidents.

You indicated that someone had unwanted access to your smartphone. If this happened more than once, please answer this and the following questions with regard to the most severe case of unwanted access.

- Who had unwanted access to your smartphone? – *Open-ended answer in text field.*
- Please briefly described what happened during this unwanted access. – *Open-ended answer in text field.*
- Please briefly describe which harmful consequences, if any, arose from this unwanted access. – *Open-ended answer in text field.*
- What good, if any, came as a result of this unwanted access? – *Open-ended answer in text field.*
- What do you think made the unwanted access possible? – *Open-ended answer in text field.*

B. ONLINE-SURVEY CODEPLAN

B.1 Reasons for Using Code Lock

1. Protect from specific attacker
   (a) Coworker
   (b) Spouse
   (c) Roommate
   (d) Own children
   (e) Other *unwanted* individual/Stranger
   (f) Unspecified people
   (g) Friends
2. Protect information
   (a) In general/entire phone
   (b) Private/personal/sensitive information
   (c) Generally *confidential* information
   (d) (Confidential) *Work* info
   (e) Emails/Messages
   (f) Photos
   (g) Contacts
   (h) Calendar
   (i) Other app-content
3. Protect from specific scenarios
   (a) Phone protected if stolen
   (b) Phone protected if lost
   (c) Phone protected if misplaced
   (d) Phone protected if left unattended
   (e) Someone casually picking up the phone
   (f) Unwanted disclosure, Pranks
   (g) "Messing up" the phone
4. Protect certain action
   (a) Calls
   (b) Internet use
   (c) Using services
   (d) Play with phone
   (e) Deletion
   (f) Accidental input
   (g) Accidental calls
   (h) Other accidental use
   (i) Stealing data
5. Lock is mandatory
   (a) Forced by employer
   (b) Forced because of custom certificate
6. Context
   (a) Work
   (b) Sleep
   (c) Death
7. Given protection goal
   (a) Increase difficulty of access
   (b) Increase time to recover/find phone
   (c) Access control
   (d) "Safety"/Security
   (e) Privacy
   (f) Encrypt data
8. Other
   (a) Set by default
   (b) Having a lock is a habit
   (c) Allows second wallpaper
   (d) Previous bad experience
   (e) Peace of mind
   (f) Don't know
   (g) Curiosity
   (h) Used to Locking
9. Off Topic/Other
10. "Protection", Unspecific/general

B.2 Reasons for Not Using Code Lock

1. Inconvenience
   (a) It's a hassle/annoying/easier without
   (b) Mental burden
   (c) Takes too much time/want instantly available
   (d) Use it too frequently
   (e) Don't feel like it/Just don't like it
   (f) Too impatient
   (g) Not eyes-free
   (h) Used to existing system
2. Dislike
   (a) Passwords
   (b) Unlocking in general
3. No threat
   (a) General: Don't need security/not concerned about security
   (b) Nothing to hide/not worried about privacy
   (c) No sensitive data on phone
   (d) Not afraid of loosing phone
   (e) Keep physically secured/never leave unattended
   (f) Trust people around me/no one who wants to access
   (g) Use only in private environment
   (h) Phone not valuable
   (i) No bad experiences so far

4. Locking may cause problems
   (a) May forget my password/PIN/pattern
   (b) Child may lock parent out of own phone
   (c) Want finder to be able to contact me
   (d) Phone accessible in emergency
   (e) Shared use
5. No specific reason/Carelessness
   (a) Didn't consider it/think about it
   (b) Haven't gotten around to set it up yet
   (c) Don't care
   (d) Don't know how to set it up
   (e) Don't know if available
   (f) Laziness
6. Technical Reasons
   (a) Phone doesn't support lock (sic)
   (b) Broken Screen
   (c) Slows down phone
7. Protect phone using another measure
   (a) Use locking only in specific situations
   (b) Rely on remote locking
   (c) Leave phone at home
   (d) App-specific lock
8. Rightful punishment
9. Off topic/other
10. No protection possible/is not secure anyway

B.3 Situations

These codes were attached to statements in which participants mentioned where they take extra measures.

1. Public spaces
   (a) "Out", General public space
   (b) Events (Sport, Concert)
   (c) Airport
   (d) Public transport (plane, train, bus)
2. Semi-Public Spaces
   (a) Gym/Sports/Workout/exercise
   (b) Party/Club/Bar
   (c) Work/School
   (d) Shopping
   (e) Restaurant
   (f) Cinema
3. Private spaces
   (a) Home
   (b) Car
4. Unknown Spaces
   (a) Travel/Vacation
   (b) Unfamiliar places
5. (Hardware-)Risky Conditions
   (a) Water (Swimming, Boat, Rain)
   (b) Sports
   (c) Dirt (Beach, Cooking, Mow the lawn)
   (d) Jail
   (e) Lifting objects
6. Crowds
   (a) General crowded places
   (b) High foot-traffic area
7. Clothing
   (a) No Pockets
   (b) Other
8. Persons
   (a) Suspicious/nosy persons
   (b) Unknown/Untrusted persons
   (c) Family, Kids
   (d) Ex-Partner
   (e) Coworkers/Other pupils
   (f) General other people
   (g) Friends
   (h) Partner (girlfriend, boyfriend, spouse)
9. Uncontrolled Situations
   (a) General less cautious situation
   (b) General unattended
   (c) Left charging
   (d) Drinking/Socializing
   (e) Sleeping
   (f) Checked bags/Airport Security
10. Discomforting Environment
   (a) Night/badly lit places
   (b) Dangerous neighborhood/somewhere sketchy
11. Device sharing
12. Data
   (a) Inappropriate
   (b) Sensitive
13. Long idle times
14. Not at home
15. Activity
   (a) Walking
   (b) Quick errand
   (c) Exercising
   (d) Lodging/overnight stay
16. Off Topic/Other

B.4 Extra Measures

These codes were attached to statements in which participants mentioned which additional measures they take.

1. Safer mobile storage
   (a) Wear close to body (e.g. in pocket)/keep out of sight
   (b) Pocket in handbag/hide in purse
   (c) Zippered pocket
   (d) Inside pocket
   (e) Backpack
   (f) Have someone else carry it
   (g) Keep in hand
   (h) Strapped to belt/hip

2. Safer static storage
   (a) At home
   (b) Leave/hide in car (e.g. glove box)
   (c) Locker/Drawer
   (d) Leave in hotel safe
   (e) Pocket instead of purse
   (f) Never leave in car
   (g) Other/general
3. Technical Measures
   (a) Turn off
   (b) Enable lock screen
   (c) Have remote wiping/find my phone enabled
   (d) Encrypt data
   (e) Remove memory card
   (f) Extra protection for specific apps
   (g) Disallow access to specific apps
   (h) Mute it
   (i) Remove battery
   (j) Have backup
   (k) Use biometrics
4. Pay extra attention
   (a) Check repeatedly if phone is still there/ Monitoring phone (alerts)
   (b) Use it less/minimize interaction
   (c) Monitoring bystanders
   (d) Don't leave unattended
5. Physical measures
   (a) Sturdy/special case
   (b) Protect from water
   (c) Leave on highest shelf (kids)
   (d) Screen protector
   (e) Micro-cloth
   (f) Don't give to others
   (g) Other physical measure
6. Data
   (a) No sensitive data
   (b) Different accounts
7. General/other safe place

C. MINI-QUESTIONNAIRES

Participants were randomly presented with one of two mini-questionnaires. One concerned risks arising during unlocking and the other concerned risks to the data on the phone in general.

C.1 Unlocking Questionnaire

1. Who has a view on the contents of your screen right now?
   (a) Unknown Person
   (b) Known Person
   (c) Nobody
2. IF NOT (1) NOBODY: Please rate how likely it is that someone is watching your screen right now.
   (a) 5-point numeric scale ("very unlikely" to "very likely")
3. IF NOT (1) NOBODY: Please rate how severe it would be if this person was watching your screen right now.
   (a) 5-point numeric scale ("not severe at all" to "very severe")
4. WITH CODE LOCK: Did you try to protect your code input?
   (a) Yes/No
5. WITH CODE LOCK: Would you rather not have had a code lock in this situation?
   WITHOUT CODE LOCK: Would you rather have had a code lock in this situation?
   (a) 5-point numeric scale ("do not agree" to "agree")
6. In what kind of environment are you right now?
   (a) Private
   (b) Semi-Public
   (c) Public
7. How sensitive is the data you are going to access now?
   (a) 5-point numeric scale ("not sensitive at all" to "very sensitive")

C.2 Data Risk Questionnaire

1. Please rate this unlock.
   (a) 5-point numeric scale ("not annoying at all" to "very annoying")
2. Did you take any additional measures to protect your phone since last using your phone?
   (a) Hidden in clothes/purse
   (b) Left in a safe place
   (c) Other: <Text>
3. Could someone have had unwanted access to your phone since you last used it?
   (a) Yes/No
4. IF YES (3): Who could have had unwanted access?
   (a) Unknown Person
   (b) Known Person
5. IF YES (3): How likely do you think it is that this person actually did access the device?
   (a) 5-point numeric scale ("very unlikely" to "very likely")
6. IF YES (3): How severe would the consequences of this access be, had it actually happened?
   (a) 5-point numeric scale ("not severe" to "very severe")
7. In what kind of environment has the phone been since you last used it?
   (a) Private
   (b) Semi-Public
   (c) Public

D. SAMPLING OVERVIEW

The histograms in Figure 8 provide an overview of all participants' aggregated use (bottom facet) by time of day during the experiment (27 days). The top facet shows the corresponding number of mini-questionnaires of both types participants completed.

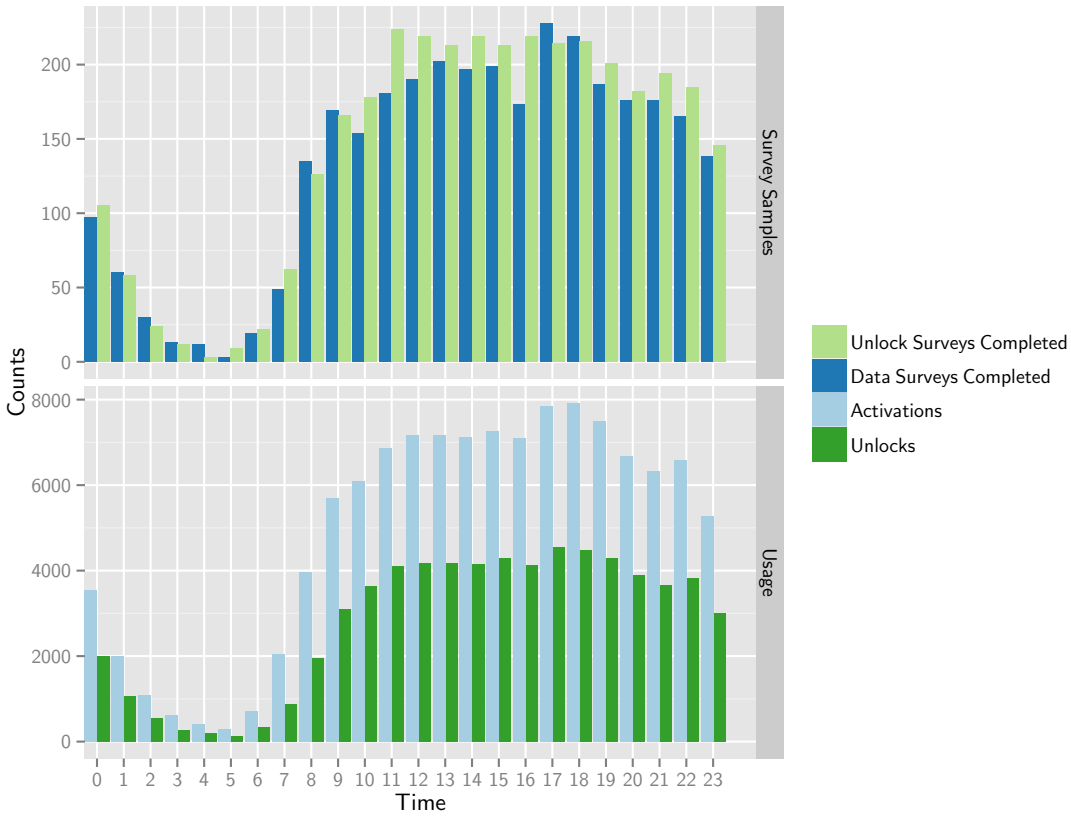


Figure 8: Overview of sampled situations per time of day, comprising number of mini-questionnaires shown as well as cumulative number of activations and unlocks.

Applying Psychometrics to Measure User Comfort when Constructing a Strong Password

S M Taiabul Haque[#] Shannon Scielzo[*] Matthew Wright[#]
eresh03@gmail.com, scielzo@uta.edu, mwright@cse.uta.edu
[#]Department of Computer Science and Engineering
[*]Department of Psychology
University of Texas at Arlington, USA

ABSTRACT

As mobile devices become increasingly common for accessing services online, the security of these services in turn depends more on password entry on these devices. Unfortunately, users are not comfortable with existing textual password entry mechanisms on mobile phone handsets. In this study, we investigate this issue of user comfort from the viewpoint of psychometrics. By applying standard techniques of psychometrics, we develop a questionnaire (known as a scale in psychometrics) that measures the comfort of constructing a strong password when using a particular interface. We establish the essential psychometric properties (reliability and validity) of this scale and demonstrate how the scale can be used to profile password construction interfaces of popular smartphone handsets. We also theoretically conceptualize user comfort across different dimensions and use confirmatory factor analysis to verify our theory. Finally, we highlight several issues related to scale development and discuss how psychometric approaches may be useful in general for measuring various subjective concepts that are related to usable security.

Categories and Subject Descriptors

Security and Privacy [**Human and Societal Aspects of Security and Privacy**]: Usability in Security and Privacy; Human-centered Computing [**Human Computer Interaction (HCI)**]: Empirical Studies in HCI

General Terms

Security; Measurement; Human Factors

Keywords

Psychometrics; Questionnaire; User Comfort; Mobile Password Entry

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

1. INTRODUCTION

Password entry on input-constrained devices such as mobile phone handsets is fraught with usability problems. Jakobsson et al. report that the time-consuming and error-prone operation of password entry annoys users more than lack of coverage, small screen size, or poor voice quality [19].

This poor user experience clearly undermines the usability of sensitive security systems that are developed for mobile platforms (mobile banking, for example). According to Whitten and Tygar in their seminal paper, a security system is deemed to be usable if "people are sufficiently comfortable with the interface to continue using it" [43]. In a mobile banking system, a user is required to type her entire password by using the mobile phone keypad (i.e. no "remember me" option) each time she intends to log in to her bank account. Thus, user frustrations over password entry on mobile handsets could undermine the usability of mobile banking as a whole, as well as other security systems on mobile devices.

Since "frustration" and "comfort" are subjective psychological concepts, it is not a straightforward task to measure the level of comfort a user feels when using the interface of a security system. According to psychology researchers [32, 28, 41], merely asking "How comfortable are you with the interface of this security system?" is not enough in this case for three reasons. First, a single question lacks scope to represent a complex psychological concept such as comfort. Just as a single question can not measure intelligence, a single question is not sufficient for measuring one's level of comfort. Second, a single question can only categorize people into a small number of groups, thus limiting the ability to finely discriminate levels of comfort. Third, any individual question has a considerable amount of measurement error associated with it. When multiple questions are asked and the response scores are summed to get a total score, this error tends to average out.

For these reasons, to measure complex psychological concepts such as "frustration" or "comfort", psychology researchers develop a set of questions that meets some widely agreed upon specific statistical criteria. In fact, a separate branch of psychology has evolved in this regard, which is known as *psychometrics*. Psychometrics concentrates on developing and validating questionnaires or tests that are used for assessing knowledge, attitudes, abilities, or personality traits.

In this work, we adopt the methods of psychometrics to develop a questionnaire, also called a *scale*, for measuring the comfort of constructing a strong password when using a particular interface. We first use expert opinions to guide the

creation and selection of questions and then assess our questionnaire for reliability and validity, the two essential psychometric properties of a scale. To this end, we conducted two user studies where we administered the questionnaire to undergraduate students from different majors and analyzed their responses. We find that our questionnaire meets all of the requirements for reliability and validity for a psychometric scale: it is consistent, complete, accurately focused, and capable of predicting certain real-world outcomes.

Through a separate user study, we evaluated the password construction interfaces of popular smartphone handsets by using our scale, where the interface of iPhone was rated the most comfortable by the participants. The results of these studies demonstrate that our scale can be used effectively to measure user comfort during a password entry operation on a mobile handset.

Based on certain observations, we further shorten our scale while maintaining the diversity of interface quality evaluation. We hypothesize a specific theory about user comfort in constructing a strong password by conceptualizing comfort across several factors and build a four-factor model. This model is helpful to explain why a particular interface is more comfortable to use than another one. We employ confirmatory factor analysis, a widely used statistical method in psychometrics, and find that our collected user responses fit the model we developed.

To the best of our knowledge, despite being a well-developed field, psychometrics has not been applied in usable security to develop and evaluate questionnaires. We believe that our work paves the way for applying the techniques of psychometrics in measuring various subjective concepts that are associated with usable security. In particular, our psychometric approach of measuring comfort can be generalized to measure whether people are sufficiently comfortable with other security-related interfaces, such as anti-virus systems, personal firewalls, privacy tools for the Web, and encryption software. This, in turn, would be helpful to understand in what ways that interface is usable or not, according to the working definition of usable security as provided by Whitten and Tygar [43].

The paper is organized as follows. In Section 2, we briefly discuss psychometrics and highlight related works. We describe all the steps of our scale development effort in Section 3. Section 4 illustrates how using our scale can profile password construction interfaces of different handsets. We present our factor analysis results in Section 5. We discuss several issues related to scale development in Section 6 and shed light on future research directions in Section 7.

2. BACKGROUND AND RELATED WORK

In this section, we first briefly cover related work on password entry on mobile devices. We then provide background on psychometrics and discuss prior efforts in measurement in the fields of HCI and usable security.

2.1 Mobile Device Password Entry

Several researchers have sought a better method for entering passwords on mobile devices. By taking advantage of the auto-correction and auto-completion features of mobile handsets, Jakobsson and Akavipat implement a novel password-entry method called *fastword* [18]. Haque et al. propose a modified keyboard layout for inserting digits and special characters [14]. Despite a few limitations, both methods effectively help users to construct stronger passwords on mobile handsets [18, 14]. Other researchers have evaluated the usability of graphical password schemes on smartphones [3, 38].

The basic motivation for all these works on password entry on mobile devices stems from the realization that users are not comfortable with the existing textual password construction interfaces. In our current work, we attempt to systematically measure this user comfort in various dimensions.

2.2 Psychometrics

Psychometrics is the study of measuring complex psychological concepts, or *constructs*, such as a person's motivation, anger, personality, intelligence, attachment, or fear [30]. Since a construct is not a concrete material in the visible world, measuring a construct is not a straightforward task. For example, we know how anger looks, but we cannot describe in meters or grams how much anger a person feels. Psychometrics provides guidance to systematically develop and test a scale to measure this kind of psychological construct. In psychometrics, the basic component of a scale is referred to as an *item*. Items can be questions, true-false statements, or rating scales.

Although the field of psychometrics has been developed for measuring psychological constructs, we observe that the techniques of psychometrics may be suitable for other abstract constructs that concern human feelings and performance. The core function of psychometrics is to assign numbers to observations in a way that best allows people to summarize the observations. In other words, it tries to measure the psychological construct in a meaningful and interpretable way. Since usability is also an abstract construct [29], we believe that the techniques of psychometrics would be helpful in measuring the usability features of a security system in a meaningful and interpretable way.

2.2.1 Reliability and Validity

Let C be an arbitrary construct, such as happiness. At any given point in time, a person has a true level of happiness, namely X_T. A psychometric scale developed for measuring happiness, if administered on that person, will produce an observed level of happiness, namely X_O. The core job of a psychometrician is to develop a scale that produces a score X_O that approximates X_T as closely as possible.

The relationship between X_T and X_O can be formulated in this way [5]:

$$X_T = X_O + X_S + X_R, \tag{1}$$

where X_S comes from systematic sources of error and X_R comes from random sources of error. X_S refers to errors resulting from underlying stable characteristics of the construct, while X_R refers to errors that result from transient personal factors.

A scale may be characterized by two properties: reliability and validity. Reliability is the degree to which a scale produces stable and consistent results, and high reliability is indicated by low values of X_R (low random error). For example, if a person measures the weight of a penny several times by the same scale and always receives the same result, the scale is reliable. Validity is the degree to which a scale measures what it is purported to measure, and high validity

is indicated by low values of both X_R and X_S (both low random and systematic errors). Note that a reliable scale may not be valid, such as if the scale consistently indicates that the penny weighs 100 kg., it suffers from high systematic error.

Although there are many validity classifications, one of the most prevalent frameworks recommends assessing validity from three perspectives: *content validity*, *construct validity*, and *criterion-related validity* [31, 13, 1]. Content validity refers to the extent that a scale represents a given construct, i.e. the extent to which the content domain of the construct is represented in its entirety, and also the extent that items in the scale only represent the construct of interest. Construct validity refers to the extent that a scale assesses the underlying construct it is supposed to assess, i.e. whether the scale is accurately focused. Criterion-related validity, on the other hand, is the degree to which a scale score predicts meaningful outcomes in a real-life situation.

A sound psychometric scale should be reliable and valid in all three ways to have any meaningful application. As pointed out by Nunnally, however, validity is not an all-or-nothing property, rather it is a matter of degree [31].

2.2.2 Framework

The appropriate steps for developing a scale ultimately depend on the construct. Psychometricians have to know a number of tools and methodologies and have a thorough understanding of the construct to be measured so as to find the best mechanisms for developing and assessing the efficacy of the intended scale. There is substantial debate in the field in regards to the specific steps to employ to ensure the highest levels of validity. For example, marketing researchers often focus solely on differences due to stimuli changes, whereas psychology researchers are oftentimes interested in individual differences [5, 32]. However, there appears to be a consensus that comprehensive efforts that employ techniques from numerous perspectives are the most effective. We thus sought recommendations from various sources and applied heuristics from both marketing and psychology perspectives.

Our work is primarily based on the approach outlined by Nunnally in his various books [30, 31, 32][1], but we also considered the recommendations provided by other notable psychometricians and statisticians, including Churchill [5], Parasuraman [33], and Kaiser [21].

2.3 Psychometrics in HCI

In the existing literature on usable security, we have not found any instance of applying psychometrics to address a particular usability issue. HCI researchers, however, have adopted psychometric approaches to measure user satisfaction. Usability questionnaires such as SUMI (Software Usability Measurement Inventory) [23], QUIS (Questionnaire for User Interaction Satisfaction) [4], and MPUQ (Mobile Phone Usability Questionnaire) [36] were developed by following psychometric approaches. Sauro and Lewis employed factor analysis, a statistical method widely used in psychometrics, to identify the underlying factors or dimensions of usability [37].

In another work, Lewis evaluated the psychometric properties of four existing IBM questionnaires that were developed for measuring user satisfaction with computer system usability [27]. He provided the questionnaires to different users after they had completed certain computer tasks and asked them to express their opinion about the computer system they had just interacted with. By analyzing the response scores and measuring the reliability and validity, he concluded that all the questionnaires have acceptable psychometric properties, thus allowing usability practitioners to use them with confidence for measuring user satisfaction with different computer systems.

2.3.1 Psychological Approach in Usable Security

Our effort to apply techniques from psychometrics to address a usability problem is inspired from the observation that psychological approaches have been helpful to solve other usable security problems. A notable example of this is the work of Jaferian et al., who applied *activity theory*, a revolutionary theory originating in Soviet psychology [26], to develop a set of heuristics for evaluating the usability of IT security management tools [17]. Their results demonstrated that the heuristics performed well in identifying usability problems.

3. SCALE DEVELOPMENT STEPS

We now describe all the steps of our scale development procedure. We use the terms "layout" and "interface" interchangeably in the remainder of this paper. For performing some of the statistical calculations, we used R packages such as *psych* and *nFactors*.

3.1 Domain Specification and Initial Item Pool Generation

The first step in developing a scale for measuring a construct is to specify the domain of the construct. A researcher must understand the construct thoroughly and determine its scope: what to be included and what to be excluded [5]. For example, in the context of our current work, "comfort of constructing a password" and "comfort of constructing a strong password" are two different constructs. The former mainly refers to general typing experience and may undervalue issues like "How easy is it to insert a special character in this layout?", which is an important consideration for a strong password.

Churchill recommends performing a literature search and an experience survey for specifying the domain of the construct and generating the initial item pool [5]. An experience survey involves consulting a group of people who are considered to be knowledgeable in the domain. We conducted such a survey by forming a panel of two password researchers and two mobile UI specialists. We also consulted with expert researchers from marketing and psychology to obtain more substantive insights about the scale development procedures. A one-on-one session was held with each of the panel members.

The marketing expert recommended to review the existing scales that have been developed to measure user engagement or customer satisfaction for various activities performed with a computer or mobile phone (online shopping, for example). The psychology expert suggested that we consider emotional

[1] Nunnally's seminal book "Psychometric theory", published in 1967, had been widely used as the primary textbook in basic psychometric courses [30]. Eleven years later, he published a second edition by incorporating the new ideas that had been introduced over the decade [31]. He also co-authored a book with Bernstein, a notable clinical psychologist [32].

or cognitive hindrances such as frustration or confusion that might affect the password construction activity. The mobile UI specialists recommended that we consider subtle typing issues, such as key sensitivity and inter-key distance, which are associated with the user experience when typing on a particular layout. The password researchers focused more on entropy and were interested to observe how different keypad layouts would affect the frequency of using capital letters, digits, and special characters when constructing a new password by using those layouts.

After consulting with all the panel members and reviewing the relevant literature, we generated an initial pool of 32 items.

The first set of items was developed to assess the ease of using a specific layout to construct a strong password. The strength of a password is associated with its length and the frequency of uppercase letters, digits, and special characters (see Section 6.2 for more discussion about password strength). Items in this category directly focused on assessing how easily a user could type an uppercase letter and insert digits and special characters by using a specific layout. We also conjecture that a user would be motivated to type a longer password if her general typing experience is good when using a specific layout. Thus, we tried to capture the general typing experience of a user through this set of items. Accordingly, items in this category focused on issues like ease of editing, key sensitivity, and inter-key distance.

1. It was easy to type an uppercase letter in this layout.
2. It was easy to insert a numeric digit in this layout.
3. It was easy to insert a special character in this layout.
4. It was easy, overall, to type passwords using this layout.
5. I could easily type the exact letter that I wanted to type in this layout.
6. The distance between the keys was not very close in this layout.
7. The keypad of this layout was too much sensitive to my touch.
8. The keypad of this layout was too little sensitive to my touch.
9. It was easy to make edits when typing in this layout.
10. The keys were marked with familiar symbols in this layout.
11. I could clearly see the keys in this layout.
12. It was easy to type using both hands in this layout.

As pointed out by the psychology expert, emotional and cognitive hindrances might adversely affect the password construction activity of a user. The second set of items reflected this direction and were written as reverse-coded items [28]. Consequently, the wording of the items reflected negative connotations such as "annoyance", "error", "confusion", and "restriction".

13. I felt annoyed when typing an uppercase letter in this layout.
14. I felt annoyed when inserting a numeric digit in this layout.
15. I felt annoyed when inserting a special character in this layout.
16. I felt frustrated, overall, when typing passwords using this layout.
17. I made more errors in this layout when typing.
18. It was confusing trying to find some keys in this layout.
19. I found this layout confusing to use when I was typing an uppercase letter.
20. I found this layout confusing to use when I was inserting a numeric digit.
21. I found this layout confusing to use when I was inserting a special character.
22. The current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords.
23. The current method of inserting a numeric digit in this layout restricted me from using more digits in my passwords.
24. The current method of inserting a special character in this layout restricted me from using more special characters in my passwords.
25. The current method of typing in this layout restricted me from typing a longer password.

The final set of items targeted user satisfaction. Items in this category addressed whether the users felt that there should be an easier way to insert digits or special characters, whether they were able to type quickly by using the layout, and so on.

26. I want an easier method of typing an uppercase letter in this layout.
27. I want an easier method of inserting a numeric digit in this layout.
28. I want an easier method of inserting a special character in this layout.
29. I was able to quickly type an uppercase letter in this layout.
30. I was able to quickly insert a numeric digit in this layout.
31. I was able to quickly insert a special character in this layout.
32. I was able to quickly type passwords using this layout.

3.2 Content Validity Assessment

After generating the initial item pool, the items were subjected to an assessment of content validity. As mentioned before, content validity refers to the extent to which a scale represents the content domain of a construct [12]. Content validity should be assessed immediately after developing the items, as this provides an opportunity to refine the items before making large investments in administering the items to a sample population [39, 35].

We assessed the content validity of each item by following Lawshe's guidelines [25]. Lawshe proposes forming a panel of subject matter experts and asking each of them to rate each item in terms of whether the knowledge or skills measured by that item is "essential", "useful, but not essential", or "not necessary" to the performance of measuring the construct. He developed a formula for measuring the content validity of each item [25]:

$$CVR = \frac{(n_e - \frac{N}{2})}{\frac{N}{2}}, \tag{2}$$

where CVR stands for *content validity ratio*, n_e is the number of panelists indicating that the item is "essential", and N is the total number of panelists. Lawshe also provides a table of critical values of CVR for a given size of subject matter expert panel [25]. According to his recommendation,

an item can be retained if its CVR value exceeds the critical value. Accordingly, we formed a panel of eight subject matter experts and asked them to evaluate our initial set of items. We recruited mobile application developers with at least two years of experience in working with mobile UI as our subject matter experts. They were explained beforehand about the purpose of the scale and the association between a strong password and a particular layout.

Out of 32 items, 19 items were retained (see Table 1 for the list of retained items), as their CVR was higher than the 0.75 threshold recommended in Lawshe's table for a panel of eight subject matter experts. Two psychometricians reviewed the wordings of the retained items to avoid ambiguity.

3.3 Initial Scale Administration – Study 1

Using the retained items, we then conducted a laboratory study for the purpose of testing the psychometric properties of the selected items. Specifically, the study was designed to not only collect responses from participants to examine their patterns, but to also examine whether participants' responses would change systematically in response to changes in the stimuli (the interface) being rated.

The study was administered through the *research pool* of the Department of Psychology of the University of Texas at Arlington (UTA). The pool is used to assign partial course credits to students taking an introductory course in psychology and extra credit for some advanced elective courses. Any study conducted through the pool can draw a diverse set of participants, because most of these courses are offered to majors from all departments of the university.

Researchers who collaborate with the Department of Psychology can post a brief description about their studies to the pool. Students in the research pool can view all the studies and sign up for those that interest them.

Participants.

A total of 49 undergraduate students (28 female and 21 male) signed up and participated in our study for course credit. Written informed consent was obtained from each participant.

Material.

Three layouts were used as the conditions for the study: (a) mobile phone with touchscreen keypad layout, (b) mobile phone with physical keyboard layout, and (c) computer keyboard layout. We used a within-group experimental model where each participant used all the three layouts to construct passwords.

For this study, we used a Motorola MILESTONE A853 mobile handset running Android 2.1. This handset features both a QWERTY-type touchscreen keypad and a slide-out physical keyboard. Each participant was asked to construct passwords by using both of these layouts and also a standard desktop computer keyboard.

Procedure.

First, we asked each participant to construct new passwords by using the one layout for two banking websites: Chase.com and Wellsfargo.com. We wanted the participants to construct long passwords that would contain uppercase letters, digits, and special characters. To protect their security, we explicitly told the participants not to provide any of their existing passwords or any of the passwords they had previously used. For Chase.com, the participants were presented with the following scenario:

> "Chase is one of the largest banks in the US and it has an ATM on campus. Imagine that you are creating an account at Chase.com for online banking. You have reached the final step of creating your new account and you need to create a strong password (a password that is long and contains uppercase and lowercase letters, digits, and special characters). Proceed to the next page to input your new password. Do not provide a password that you currently use or have previously used for any accounts. Also, do not use any confidential or personally identifiable information in your password."

When they clicked OK, the password construction page appeared. Once they constructed the password for Chase.com, a similar scenario was presented for Wellsfargo.com.

Next, the participants were asked to type five fixed passwords. These fixed passwords were from seven to thirteen characters long and contained multiple uppercase letters, digits, and special characters (TRoub@dor!123, for example).

After a participant finished typing the fixed passwords, she was asked to evaluate the layout by using the 19-item scale. The items were randomly ordered to avoid any ordering effects. A 5-point Likert scale (anchored by 1 = "strongly disagree", 5 = "strongly agree") was used to capture the participants' responses. We note the difference between a Likert-type item and a Likert scale. A Likert-type item is a single question or statement and it falls into the category of ordinal level data. A Likert scale, on the other hand, is composed of multiple Likert-type items. The responses for the individual items are combined and then averaged to obtain a final scale score. Likert scale data are analyzed at the interval measurement scale and descriptive statistics like mean/standard deviation and statistical methods like ANOVA could be used in this regard [15].

The same process was then repeated for the second and third layouts. The order of the layouts was randomized for each participant.

Overall, each participant typed seven passwords in each layout. Out of these seven passwords, two were selected by the participant and five were given by us. The only reason for asking them to construct two of their own passwords was to ensure that they would be able to properly respond to the four items related to "restriction" (items 9-12 in Table 1). When administering the scale to the participants, we also modified these items slightly to emphasize new password construction. For example, item 9 was written in this way "When I was constructing a new passsword for the two banking websites, the current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords".

We note that we did not use deception in this study; the participants were directly asked to construct and type passwords. We also did not store any of their passwords. Given the nature of the scale and the relative lack of consequences (e.g. no embarrassment, no reason for responding dishonestly), there was no reason for hiding the true intent of the study at this stage of scale development. Similarly, participants were free to provide suggestions or concerns regarding

Table 1: Reliability Analysis. Cronbach's α value is 0.96.

Item	Corrected Item-total Correlation
1. It was easy to type an uppercase letter in this layout.	0.76
2. It was easy to insert a numeric digit in this layout.	0.80
3. It was easy to insert a special character in this layout.	0.83
4. It was easy, overall, to type passwords using this layout.	0.90
5. I felt annoyed when typing an uppersace letter in this layout.	0.73
6. I felt annoyed when inserting a digit in this layout.	0.77
7. I felt annoyed when inserting a special character in this layout.	0.75
8. I felt frustrated, overall, when typing passwords using this layout.	0.83
9. The current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords.	0.77
10. The current method of inserting a numeric digit in this layout restricted me from using more digits in my passwords.	0.78
11. The current method of inserting a special character in this layout restricted me from using more special characters in my passwords.	0.75
12. The current method of typing in this layout restricted me from typing a longer password.	0.78
13. I could easily type the exact letter that I wanted to type in this layout.	0.75
14. It was easy to make edits when typing in this layout.	0.75
15. It was easy to type using both hands in this layout.	0.62
16. I was able to quickly type an uppercase letter in this layout.	0.77
17. I was able to quickly insert a numeric digit in this layout.	0.82
18. I was able to quickly insert a special character in this layout.	0.81
19. I was able to quickly type passwords using this layout.	0.89

the items and the layouts. Upon completion of the required tasks for each condition, participants were asked to evaluate their experience by using the item list.

As each participant evaluated three layouts, we collected a data set with a total of 147 evaluations. The scores of the reverse-coded items were inversed before adding them to the data set. There were no missing data points. We used this data set to assess the reliability and the validity of our 19-item scale.

3.4 Reliability Analysis

We first assessed the reliability of our scale. Nunnally points out that reliability is a necessary precondition for validity [31]. There are several types of reliability estimates: inter-rater reliability, test-retest reliability, parallel-forms reliability, and internal consistency. In his landmark paper, Churchill strongly emphasizes internal consistency over the other types of reliability [5]. For a Likert scale like ours, internal consistency is the reliability estimate that is most frequently reported [11].

Internal consistency of a scale is calculated based on the covariations between different items of that scale. It measures whether multiple items that are generated to measure the same general construct produce similar scores. For example, if a participant expresses agreement with the item "It was easy to type an uppercase letter in this layout" and disagreement with the item "I felt annoyed when typing an uppercase letter in this layout", it would indicate good internal consistency. Internal consistency can be measured statistically by calculating the Cronbach's alpha [7].

Since Cronbach's alpha usually increases as the covariations among items increase, a low Cronbach's alpha value suggests that the items are possibly not measuring the same construct. Along with the Cronbach's alpha value of the entire scale, the corrected item-total correlation values of the individual items also need to be calculated. The corrected item-total correlation value is an estimate of whether a given item is consistent with the averaged behavior of the other items. A low corrected item-total correlation value of an item would indicate that the item should be removed, as that particular item is ultimately not discriminating participants well in regards to what the remainder of the items are measuring. Nunnally recommends removing the items with corrected item-total correlation values lower than 0.30 [32]. Once these items are removed, the Cronbach's alpha should be recalculated to see whether a satisfactory value is achieved. However, if the value of Cronbach's alpha is too low, a researcher should loop back to the previous step of domain specification and item generation to find out what might have gone wrong [5].

The reliability results of our data set are shown in Table 1. Cronbach's alpha for the scale is 0.96, which is excellent according to the recommendation of George and Mallery [9]. This value is even arguably high, and suggests that some items could be removed and still maintain the general essence of what is being measured. We discuss this further in Section 5. Furthermore, all the corrected item-total correlation values were much larger than the cutoff value of 0.30, with the lowest correlation at 0.62. We therefore retained all the items at this point.

3.5 Construct Validity Assessment

We assessed the construct validity of our scale through a technique called the *known-groups* method [16], which involves administering the scale to conditions/groups expected to differ due to known characteristics [34]. For example, a scale to measure the construct of "fun" should show a large difference between subjects playing a video game and subjects made to wait with nothing to do. If the conditions/groups have a significant difference between their mean scores on the scale, this provides evidence for the scale's construct validity, since this indicates that it is able to discriminate among conditions/groups that are known to be different. In other words, this indicates that the scale effectively captures the underlying construct it is supposed to capture, which is the requirement of construct validity.

As mentioned before, we asked our participants in Study 1 to construct passwords in three different conditions. In addition to two types of mobile keypad/keyboard layouts, they were also asked to construct passwords by using a computer keyboard. The computer keyboard condition was added so that we would have a known "comfortable" condition. Constructing a strong password on a computer keyboard is easier than constructing it on a mobile keypad/keyboard due to the space constraints of the mobile device and the inconvenience of capitalizing letters and inserting digits or special characters. For example, on an iPhone, one additional click is required for each shift to and from digits, and this shift presents a different keypad view to the user. On the other hand, digits can be inserted in the same way as letters on a computer keyboard.

We compared aggregated means in the two mobile conditions to the computer condition via repeated-measure ANOVA. Mean scores for the combined mobile conditions ($M = 3.32$, $SD = .79$) were significantly lower than for the computer condition ($M = 4.39$, $SD = .68$), $F(1, 47) = 90.92$, $p < .05$. This established the construct validity of our scale.

3.6 Criterion Validity Assessment – Study 2

Criterion-related validity tests the relationship between a scale score and a particular outcome. For example, in the United States, SAT scores are used to determine whether a student will be successful in undergraduate studies. Here, the criterion for success for an undergraduate student may be her first-year GPA. If her SAT score correlates positively with her first-year GPA, it would indicate that her SAT score has effectively predicted her future performance in college, thus demonstrating an evidence of the criterion-related validity of the SAT.

In our case, in order to demonstrate evidence for criterion-related validity of our scale, we selected two outcomes that are potentially related to comfort of constructing a strong password when using a particular layout:

- The length of the constructed password
- The total number of uppercase letters, digits, and special characters in the constructed password

Although there exists no empirical evidence that the comfort of constructing a strong password is related to the total number of uppercase letters, digits, and special characters, the experimental results of Haque et al. provide the primary rationale for this proposition [14]. Their results demonstrate that if users are presented with a more comfortable mobile handset interface for entering digits and some special characters, they construct passwords that contain significantly more digits and special characters [14]. As for length, we implicitly assume that the more comfort a user feels when using a particular interface, the longer her typed password would be.

In order to observe the correlation between our construct of interest and the selected outcomes, we conducted a separate study.

Participants.

A total of 30 undergraduate students (17 male and 13 female) from UTA voluntarily participated in this study, and they were recruited from a course on computer literacy. The course is offered to majors from all departments and gets a diverse set of students. In exchange for their time, students were assigned extra course credits. Written informed consent was obtained from each student, and an alternative extra credit assignment was offered to the students who were not willing to participate in our study.

Material.

In this study, participants were asked to construct passwords by using one of the two layouts of our Motorola MILESTONE A853 mobile handset (see Section 3.3). Each participant was randomly assigned to one of the layouts to construct passwords. Since we collected the passwords of the participants for this study and analyzed them, we used deception in this study so that the participants would construct passwords just the way they do in real-life situations.

Procedure.

We designed this study so that it appeared to the participants as if they were opening a new bank account at Chase.com. They were asked to complete a set of tasks that resembled the usual steps of creating a new online bank account. Password construction was framed as one of these multiple tasks, not as the primary task.

The participants were first presented with the following instructions:

> "Chase is one of the largest banks in the US and it has an ATM on campus. Imagine that you are creating an account at Chase.com for online banking. Proceed to the next page to start creating your new bank account."

When a participant clicked OK, she was asked to enter dummy values given to her on a piece of paper for the following fields: name (containing both uppercase and lowercase letters), account number, address (containing multiple special characters), phone number, and email address. These tasks ensured that the participants were familiar with the

typing interface, including entering uppercase letters, digits, and special characters that would be needed for a strong password. Next the participant was asked to answer a few questions like "Do you want overdraft protection for your new account?" and "How much daily withdrawal limit do you want?". Finally, the participant was redirected to the password construction page where she was asked to construct a strong password for the new account. We did not enforce any requirement for length or the use of uppercase letters, digits, or special characters, though we did offer a hint for what a strong password is in our instructions:

> "Please create a strong password (a password that is long and contains uppercase letter and lowercase letter, digit, and special character) for your new account. Proceed to the next page to input your new password. Do not provide a password that you currently use or have previously used for any accounts. Also, do not use any confidential or personally identifiable information in your password."

After a participant finished all these steps, she was asked to evaluate the password construction experience on her assigned layout by using our 19-item scale. As with Study 1, the items were randomized and a five-point Likert scale was used to capture the responses. For each participant, we correlated the length of the constructed password and the total number of uppercase letters, digits, and special characters with the mean score from the scale.

Mean scale score vs. length

Mean scale score and length were strongly correlated, $r(28) = .51$, $p < .05$.

Mean scale score vs. total number of uppercase letters, digits, and special characters

The correlation between mean scale score and total number of uppercase letters, digits, and special characters was moderately strong, $r(28) = .41$, $p < .05$. Furthermore, we calculated mean scale scores by considering only the 12 items that are related to uppercase letter, digit, and special character (items 1-3, 5-7, 9-11, 16-18 in Table 1), and correlated these scores with the total numbers of uppercase letters, digits, and special characters. As expected, the correlation was stronger in this case, $r(28) = .47$, $p < .05$.

According to Cohen, a validity coefficient can be interpreted in this way: less than .1 is trivial; .1 to .3 is weak; .3 to .5 is moderate; and greater than .5 is strong [6]. Based on this guideline, our correlation coefficient values were satisfactory and evident of good criterion-related validity.

4. PROFILING POPULAR SMARTPHONE HANDSET INTERFACES – STUDY 3

In order to demonstrate the practical application of our scale, we evaluated the password construction interfaces (keyboard/keypad layouts) of popular smartphone handsets through our scale. We selected three handsets: BlackBerry Curve 9300, Motorola DROID 2 A955, and iPhone 4s. The iPhone handset was selected because of its touchscreen keypad layout, while the BlackBerry and Motorola handsets were representatives of QWERTY-type keyboard and slide-out physical keyboard, respectively.

We also implemented the custom touchsreen layout proposed and designed by Haque et al. as an Android app running in a Motorola MILESTONE A853 handset [14]. It involved adding two extra on-screen rows, one containing the ten digits and the other containing ten common special characters, in addition to the default Android touchscreen keypad.

Participants.

A total of 21 undergraduate (15 female and 6 male) students from UTA participated in this study. As with Study 1, we recruited participants from the research pool of the Department of Psychology (see Section 3.3). Written informed consent was obtained from each participant.

Procedure.

Since we did not require to collect any password constructed by the participants, we used exactly the same experimental design as Study 1 for this study. Participants were asked to type two passwords of their own and five fixed passwords by using each of the four layouts (see Section 3.3). After typing the passwords by using one layout, they evaluated that particular layout by using our scale.

For each layout, we calculated the mean scale score. The iPhone 4s touchscreen keypad layout was rated the most comfortable (mean = 4.19 out of 5), while Blackberry's layout was considered the least comfortable (mean = 2.78 out of 5). The Motorola layout received a moderate score (mean = 3.32 out of 5). The custom layout of Haque et al. obtained a slightly lower score (mean = 4.13 out of 5) than that of iPhone 4s.

We note that these findings should be interpreted with caution. We discuss this in detail in Section 6.2.

5. FACTOR ANALYSIS

Factor analysis is a statistical procedure that examines the correlations or covariances among items to discover clusters of related items. In psychometrics, factor analysis is often used to identify the underlying *subconstructs* that might reside in the construct of interest. These subconstructs are also referred to as *factors*, *components*, or *dimensions*. For example, in his classic paper, Spearman uses factor analysis to posit a two-factor theory for measuring human intelligence: the general intelligence factor and the specific intelligence factor [40].

Factor analysis comprises two different perspectives: exploratory factor analytic approaches and confirmatory factor analysis. Exploratory factor analysis is used when a researcher is uncertain about the theoretical conceptualization of her construct of interest. It provides a quick way to explore the underlying factors of the construct, thus providing an opportunity to refine the theory at an early stage of scale development. Confirmatory factor analysis (CFA), on the other hand, is used when the researcher has a more specific theory about the conceptualization of the construct of interest. Based on this theory, the researcher builds a model and gathers data to examine whether the data fits the hypothesized model.

Factor analyses can provide meaningful information regarding the overarching structure of the data, and can provide guidance on how best to aggregate the data after the factor. There are a number of ways to extract factors, in-

cluding principal component analysis (PCA), principal axis factoring (PAF), maximum likelihood, and more, but PCA and PAF are most frequently used. Factor rotation is an important consideration during a factor analysis. By maximizing high item loadings and minimizing low item loadings, rotation helps to produce a more interpretable factor analysis solution. There are several rotation techniques, varimax rotation is the one that is used most commonly.

We conducted a PCA with a varimax rotation (eigenvalues greater than 1) on our data set for Study 1 [20], and it was found that items loaded on one general component. The one component accounted for 62% of the variance, and item loadings ranged from 0.61 to 0.91. PCA tends to however identify one factor, and does not allow for examination of more complicated models as is possible in CFA.

Given the strength of relations obtained across items (demonstrated via both the Cronbach's alpha and the PCA), we decided that there was undue redundancy in the items and decided to cut unnecessary items. Upon careful examination of the current items and their respective relations, we were interested to examine the extent to which a higher order factor (comfort), and four corresponding second level factors (uppercase letter, numeric digit, special character, and general typing) would fit the data based on the eight retained items.

Our hypothesis was based on the observation that issues like ease of edit ("It was easy to make edits when typing in this layout"), ability to type by using both hands ("It was easy to type using both hands in this layout"), and ability to type the exact letter that the user wants to type ("I could easily type the exact letter that I wanted to type in this layout"), in turn, result in quick and easy typing of passwords. Items 13, 14, and 15, therefore, essentially capture the quickness and easiness of general typing when using a specific layout. Furthermore, the items related to frustration and restriction actually capture the cognitive and emotional hindrances of a user when constructing a strong password by using a particular layout. Intuitively, these hindrances should prevent users from typing quickly and influence their perceptions regarding the ease of using the layout.

We therefore focused exclusively on the quickness and easiness related items and posited the following four-factor theory regarding the comfort of constructing a strong password when using a particular layout:

Factor: Uppercase letter.

1. It was easy to type an uppercase letter in this layout.
2. I was able to quickly type an uppercase letter in this layout.

Factor: Numeric digit.

1. It was easy to insert a numeric digit in this layout.
2. I was able to quickly insert a numeric digit in this layout.

Factor: Special character.

1. It was easy to insert a special character in this layout.
2. I was able to quickly insert a special character in this layout.

Factor: General typing.

1. It was easy, overall, to type passwords using this layout.
2. I was able to quickly type passwords using this layout.

We used IBM SPSS Amos (version 21) to conduct CFA and evaluate the fit of our confirmatory model. The default settings (i.e. maximum likelihood) were used, with the raw data of Study 1 supplied as an input. It was found that our proposition was supported, the data fit the overarching model well ($\chi^2(16) = 24.952$, $p = .07$, $RMSEA = .06$, $PCLOSE = .31$). Contrary to other statistical models, the null hypothesis is that the model fits the data well. Thus, a chi square value that does not reach statistical significance (i.e., $>.05$) is considered indicative of good fit. Because of the extremely conservative nature of this particular statistic (i.e. rarely do arguably good fitting models meet this criteria), RMSEA and PCLOSE statistics are also typically reported. A small RMSEA value is an indicator of good fit as a value of 0.08 or less is often considered acceptable [2]. PCLOSE is a test of statistical significance for RMSEA, with the assumption that the RMSEA= .05 (i.e. close fit). A statistically significant difference again means that the theoretical model is significantly different from the actual relationships among variables (which is not in our case, hence a good fit). Thus, the statistical results demonstrate that our model is likely a good fit for our data.

These results suggest that there appears to be four highly related factors in the scale that collectively comprise our representation of user comfort. In turn, data can be averaged to the level of the scale for most purposes (or in the case of missing data, data should be averaged at the level of factor, and then the factors averaged for representative individual indicators). Similarly, if variations in comfort based on these factors need to be examined (e.g. "Is this particular layout more comfortable for typing uppercase letters?"), then the scale can effectively do so by specifically examining those specific factor values.

6. DISCUSSION

This is the first study to date that we are aware of which specifically applies psychometric principles to develop and test a scale designed to measure how well suited keyboard or keypad layouts are in the context of password construction. We have utilized numerous frameworks and conceptualizations, and extensively tested the scale in various ways to create the most accurate, useful scale possible. From extensive content validation efforts, to examination of construct validity and analysis of factor structure, to prediction of important meaningful criteria, this scale has demonstrated very promising initial evidence.

In the subsequent sections, we first discuss the ecological validity of our study and highlight the limitations of our work. Next, we discuss several issues related to scale development.

6.1 Ecological validity

As mentioned before, Study 1 and Study 3 did not involve deception, and we did not try to hide the motive of our study from the participants for these two studies. This was in accordance with the experimental methodology for scale development studies, where users are first explicitly subjected to a certain task and later asked to evaluate the experience

by using the scale. For our case, the task was to type a few passwords (five fixed, two of the users' own) by using different layouts. The experience of constructing a password was more important here, rather than the password itself. When the users were constructing their own passwords, we involved a simple role-play scenario, since prior work has shown that it is more effective than a survey scenario in motivating the users to construct passwords more seriously [24].

For Study 2, we collected passwords constructed by the participants. Ecological validity therefore was an important consideration for this study. The results of a recent work of Fahl et al. on the ecological validity of password study reveal that passwords collected during user studies closely resemble users' actual passwords [8]. We tried our best to frame Study 2 as an experiment that asks the users to perform a real-life online task, namely creating a new online bank account by using a mobile phone handset. Password construction was one of a series of steps for completing the primary task (i.e. creating the new account), just as it would be in real life. The word "password" was not used anywhere in the informed consent document. A debriefing session was arranged at the end of the study where the deception was revealed and the participants were provided with the opportunity to withdraw their consent to participate in the study. None of the participants decided to do so and we could use all of their passwords to test the criterion-related validity of our scale.

We note, however, that our participants were not required to return on a second day to re-enter their passwords, and as such, we were not completely able to emulate the real-life password construction behavior of users.

6.2 Limitations

For this work, we quantified password strength in terms of entropy, according to the recommendation of password researchers (see Section 3.1). We do not overlook the findings of Weir et al. or Kelley et al., which demonstrate that entropy is not the most appropriate measure of password strength [42, 22]. However, since our developed scale focuses on measuring the comfort of constructing a strong password when using a particular layout, we believe that entropy is a better approximation of password strength here because it effectively captures the layout-related aspects of a strong password. Alternative measures such as guessability are more dependent on the exact password choices of users and do not clearly capture aspects related to keyboard layout, such as the use of special characters. This approximation is consistent with Haque et al. [14], a related work on password strength and keypad/keyboard layout.

For all three studies, we recruited participants from university students, who may vary considerably from other populations in their smartphone usage behavior. We plan to test our scale by using a more diverse population group in future. Ultimately, scale development is a never-ending process in which developers continually strive to understand the intricacies of the conventions in regards to any meaningful variations (e.g. does my scale predict other meaningful criteria, does it behave differently in other contexts or for other sample compositions, etc.). However, in general, this scale has demonstrated solid initial evidence of its efficacy.

Our results of Study 3 should be interpreted carefully, particularly by considering the fact that we did not control for participants' previous familiarity with the interfaces. For example, if most of the participants were iPhone users, their familiarity with the iPhone layout would probably bias them towards that layout. We conducted Study 3 for demonstrating a practical application of our developed scale, not for a definitive comparison among the interfaces.

6.3 Aggregration and application

Our shortened item list has four factors, each of which contains two items (see Section 5). Since each of the underlying factors contains the same number of items, none of the factors is underestimated or overestimated when individual item scores are combined and averaged to form a final composite score. Subsequently, depending on the intended application (examination of individual level issues with comfort, or identification of problem areas with the layout), the scale score could also be computed in terms of each factor. As a result, the scale could be used to answer more specific questions like "Which layout is more comfortable for inserting a numeric digit when constructing a strong password?" or "Which layout is more comfortable for general password typing?". This provides an additional motivation for us to conduct further experiments with this shortened scale.

6.4 Norm development

After a sound scale (reliable and reasonably valid) has been developed, depending on the intended application of the scale, the researcher should continue to conduct further experiments. If the purpose of the scale is to compare different interfaces with respect to the construct of interest, then administering the scale to different users and profiling the interfaces based on scale scores should be sufficient. Our scale can currently be used in this way.

On the other hand, if the purpose of a user comfort scale is to answer the question of whether users are sufficiently comfortable with a particular user interface, then the researcher should also develop norms for her new scale. Developing norms involves setting up standard scores for a scale. Ideally, for a 5-point Likert-type scale, mean scale scores of 3 and 4 should imply neutral and positive attitudes, respectively. However, this might not be always true. For example, a mean score of 4 might represent the highest (or lowest) score ever achieved on that particular scale. To this end, the researcher should specify the benchmark scores for her new scale. This can be done by administering the scale over a large number of users to obtain a distribution of scores and subsequently characterizing the distribution by various statistical features such as mean and standard deviation. A detailed description about the norm development procedure can be found in [10].

We believe that this norm development technique could be used to specify a standard score that would represent "sufficient user comfort" in the context of a specific security system user interface. This would be helpful to precisely find out whether users are sufficiently comfortable with a particular security system user interface, which, according to the working definition of usable security in the seminal paper of Whitten and Tygar [43], is an important consideration for measuring the usability of that security system.

6.5 Revalidation study

We note that we did not prune any items during the reliability assessment stage because all of the items had a satisfactory corrected item-total correlation value and the Cron-

bach's alpha value of the overall scale was high (see Section 3.4). If items need to be pruned at this stage, a revalidation study is recommended to be conducted with the shortened scale. This involves administering the shortened scale to a new sample which is independent to the previous sample and assessing the reliability of the shortened scale.

For our scale, however, we needed to assess the criterion-related validity by using a separate study that involved deception and collection of participants' passwords. This provided us an opportunity to reassess the reliability of our scale by using a different sample. We calculated the Cronbach's alpha for this new data set. As before (0.96), the value was high enough (0.93). This provided further evidence for the reliability of our scale.

7. CONCLUSION AND FUTURE WORK

In this work, we adopted the techniques of psychometric theory to solve a specific usable security problem: measuring user comfort when using a specific interface to construct a strong password. We followed standard psychometric theory procedures to develop a questionnaire for this purpose. This involved consulting with subject-matter experts, testing an initial set of questions with a survey, statistical analysis to refine the set of questions and validate their consistency and accuracy, and conducting a separate study to demonstrate that the questionnaire is capable of predicting certain real-world outcomes. All these results establish the two essential psychometric properties of our questionnaire: reliability and validity. Thus, the questionnaire can be used to profile the password construction interfaces of popular smartphone handsets.

Based on our observations, we further shortened our questionnaire and attempted to build a specific theory about user comfort in the context of our work. We tested this theory and the shortened questionnaire by using confirmatory factor analysis, a widely used technique in psychometric theory. The results of confirmatory factor analysis align with our theory, which encourages us to conduct further studies in future to test this shortened questionnaire. We are interested to administer this version of the questionnaire with different participants and assess its psychometric properties. If the results are satisfactory, we would replace our current questionnaire with the shortened version, since the later one is more intuitive and capable of finer-grained comparisons among different interfaces across multiple factors or dimensions.

It is likely, for example, that older individuals (such as elderly populations) may report lower comfort scores when using our scale to evaluate their password construction experience. We have some preliminary information regarding the norms for our particular samples. However, we will continue administering the scale over a diverse group of users to obtain a representative distribution of scores that is generalizable to a much broader population.

To the best of our knowledge, the current work is the first to introduce the concepts of psychometric theory in usable security. In the future, we are interested to apply psychometric theory to develop reliable, valid and conceptually meaningful questionnaire for measuring user comfort when using other security system user interfaces (antivirus or encryption software user interfaces, for example). Also, we believe that the technique of factor analysis could be helpful in identifying the underlying factors or dimensions of usability in the context of a security system. We plan to work on this in future.

8. ACKNOWLEDGMENTS

We would like to thank our panelists and the anonymous participants in our user studies. We are also grateful to Mehdi Tanzeeb Hossain for reviewing the wordings of the items. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1117866.

9. REFERENCES

[1] J. D. Brown. *Testing in language programs*. Prentice Hall Regents, Upper Saddle River, NJ, 1996.

[2] M. W. Browne and R. Cudeck. Alternative ways of assessing model fit. In *Testing structural equation models*. Sage Publications, Newbury Park, CA, 1993.

[3] H.-Y. Chiang and S. Chiasson. Improving user authentication on mobile devices: A touchscreen graphical password. In *MobileHCI*, 2013.

[4] J. P. Chin, V. A. Diehl, and K. L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *CHI*, 1988.

[5] G. A. Churchill. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1):64–73, February 1979.

[6] J. Cohen. *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum, Hillsdale, NJ, 1988.

[7] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrica*, 16(3):297–334, September 1951.

[8] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *SOUPS*, 2013.

[9] D. George and P. Mallery. *SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.)*. Allyn & Bacon, Boston, 2003.

[10] E. E. Ghiselli. *Theory of psychological measurement*. McGraw-Hill, New York, 1964.

[11] J. A. Gliem and R. R. Gliem. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. In *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, 2003.

[12] J. S. Grant and L. L. Davis. Selection and use of content experts for instrument development. *Research in Nursing & Health*, 20(3):269–274, June 1997.

[13] R. M. Guion. On Trinitarian doctrines of validity. *Professional Psychology*, 11(3):385–398, June 1980.

[14] S. M. T. Haque, M. Wright, and S. Scielzo. Passwords and interfaces: Towards creating stronger passwords by using mobile phone handsets. In *SPSM*, 2013.

[15] J. Harry N. Boone and D. A. Boone. Analyzing likert data. *Journal of Extension*, 50(2), April 2012.

[16] J. Hattie and R. W. Cooksey. Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement*, 8(3):295–305, July 1984.

[17] P. Jaferian, K. Hawkey, A. Sotirakopoulos, M. Velez-Rojas, and K. Beznosov. Heuristics for

evaluating IT security management tools. In *SOUPS*, 2011.

[18] M. Jakobsson and R. Akapivat. Rethinking passwords to adapt to constrained keyboards. In *MoST*, 2012.

[19] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *HotSec*, 2009.

[20] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrica*, 23(3):187–200, September 1958.

[21] H. F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, March 1974.

[22] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *IEEE S&P*, 2012.

[23] J. Kirakowski and M. Corbett. SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, 24(3):210–212, September 1993.

[24] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *CHI*, 2011.

[25] C. H. Lawshe. A quantitative approach to content validity. *Personnel Psychology*, 28(4):563–575, December 1975.

[26] A. N. Leontev. *Activity, Consciousness, Personality*. Prentice Hall, Englewood Cliffs, NJ, 1978.

[27] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, January 1995.

[28] J. P. McIver and E. G. Carmines. *Unidimensional scaling*. Sage, Beverly Hills, CA, 1981.

[29] N. McNamara and J. Kirakowski. Defining usability: quality of use or quality of experience? In *IPCC*, 2005.

[30] J. C. Nunnally. *Psychometric theory (1st ed.)*. McGraw-Hill, New York, 1967.

[31] J. C. Nunnally. *Psychometric theory (2nd ed.)*. McGraw-Hill, New York, 1978.

[32] J. C. Nunnally and I. H. Bernstein. *Psychometric theory (3rd ed.)*. McGraw-Hill, New York, 1994.

[33] A. Parasuraman, V. A. Zeithaml, and L. L. Berry. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1):12–40, Spring 1988.

[34] D. Pavlas, F. Jentsch, E. Salas, S. M. Fiore, and V. Sims. The play experience scale: Development and validation of a measure of play. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(2):214–225, April 2012.

[35] J. R. Rossiter. The C-OAR-SE method and why it must replace psychometrics. *European Journal of Marketing*, 45(11):1561–1588, November 2011.

[36] Y. S. Ryu and T. L. Smith-Jackson. Reliability and validity of the Mobile Phone Usability Questionnaire (MPUQ). *Journal of Usability Studies*, 2(1):39–53, November 2006.

[37] J. Sauro and J. R. Lewis. Correlations among prototypical usability metrics: Evidence for the construct of usability. In *CHI*, 2009.

[38] F. Schaub, M. Walch, B. Konings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *SOUPS*, 2013.

[39] C. A. Schriesheim, K. J. Powers, T. A. Scandura, C. C. Gardiner, and M. J. Lankau. Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, 19(2):385–417, April 1993.

[40] C. Spearman. *The abilities of man*. Macmillan, New York, 1927.

[41] P. Spector. *Summated rating scale construction*. Sage, Thousand Oaks, CA, 1992.

[42] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *CCS*, 2010.

[43] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *USENIX*, 1999.

The Password Life Cycle: User Behaviour in Managing Passwords

Elizabeth Stobert
Carleton University
Ottawa, Canada
elizabeth.stobert@carleton.ca

Robert Biddle
Carleton University
Ottawa, Canada
robert.biddle@carleton.ca

ABSTRACT

Users need to keep track of many accounts and passwords. We conducted a series of interviews to investigate how users cope with these demanding tasks, and used Grounded Theory to analyze the interview results. We found that most users cope by reusing passwords and writing them down, but with a rich variety of behaviour and diverse personalized strategies. These approaches seem to disregard security advice, but at a detailed level they involve perceptive behaviour and careful self-management of user resources. We identify a password life cycle that follows users' password behaviour and how it develops over time as users adapt to changing circumstances and demands. Users' strategies have their limitations, but we suggest they indicate a rational response to the requirements of password authentication. We suggest that instead of simply advising against such behaviour, new approaches could be designed that harness existing user behaviour while limiting negative consequences.

1. INTRODUCTION

Passwords present a difficult task for users. Users are told not to create weak passwords, not to reuse passwords on multiple accounts, and not to write their passwords down. Yet users have many passwords and are expected to create a password for every new service. Often, users are required to change their passwords at regular intervals. Taken as a whole, these requirements are difficult, if not impossible, for users to meet. In response, users develop strategies for coping as best they can. We wish to explore and understand these strategies, in the hope of identifying new ways to alleviate the difficulties.

We conducted interviews with users to find out about their coping strategies. We asked about how many accounts and passwords they have, how they create and reuse passwords, and how they handle password changes. We encouraged participants to discuss their experiences in detail, and share their motivations, fears, and password tricks.

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

Some findings were unsurprising. Users do write passwords down, and do reuse passwords. However, these are simplifications of their actual behaviour that do not tell the whole story. For example, users often write down passwords as a fallback strategy, and when they reuse passwords, they often adapt them for different accounts. We analyzed our interviews using the Grounded Theory methodology and identified some important patterns in user behaviour. We identified a "life cycle" of password use, where the user's central concern is rationing effort to best protect important accounts. Many of the specific practices are already known, and our contribution is the identification of a coherent model that highlights a consistent series of gaps between user behaviour and current tool support. We suggest that this model can inform better ways to support users in their behaviour, rather than providing unrealistic guidance.

In the following section, we outline related work. We then describe our methodology and the details of our interviews. Section 4 presents an overview of the results, and Section 5 documents the step-by-step process of our qualitative analysis. We then suggest some implications of our findings, and our conclusions.

2. BACKGROUND

Alternatives to passwords exist in the form of biometrics and security tokens, but these have issues with privacy, theft, and the huge infrastructural costs of deployment and maintenance.

Deployed solutions to the password problem consist mainly of password managers, which store and enter users' passwords, thus saving the user from remembering their passwords or which passwords are associated with which accounts. Browser-based password managers save passwords when they are typed into the appropriate fields, and then automatically input them when the page is visited again (often without authentication). Dedicated password managers (such as LastPass [14]) typically work in one of two ways [7]: they either generate a password at login by hashing the user's master password together with information from the website, or they store the user's passwords in a password "wallet" which is protected by a master password (which may be required at every login).

Existing research on password managers has shown that they can have usability problems that affect their ability to securely manage users' passwords. A study of two password managers found that both managers had significant usability issues [7]. Worse, participants had poor mental models for how the software worked, and these poor mental models led

them to make dangerous and unrecoverable security errors.

Another solution to the password problem is single sign-on, where one party authenticates users for multiple websites. At login, the user presents their credentials to the authenticating party, who checks the credentials and relays the results to the website. Examples of single sign-on entities are Facebook, Google, and OpenID. A study of OpenID [20] found that adoption was hindered because it did not fit into users' existing password management techniques, and users were concerned about trusting a single entity to login to multiple sites.

A mismatch between security expectations and users' abilities has long been identified, and users develop coping strategies as a response. These disconnects can lead to the misuse or avoidance of security mechanisms [2].

One coping strategy for having multiple passwords is to reuse passwords across multiple accounts. This strategy is widely employed by users [10, 9, 11], but has security risks. If a reused password is discovered (e.g., through a leaked password set), an attacker may be able to gain access to several accounts. Das et al. [8] found that 43% of all passwords in their data set were reused across multiple accounts, and showed that password reuse can be leveraged for more efficient password attacks.

Several studies have investigated the number of passwords and accounts that users possess. Gaw & Felten [10] found that undergraduates had an average of about 12 accounts, but they had fewer unique passwords and password reuse was rampant. The study also found that most participants cited easier memorability as their reason for password reuse, and that participants classified their accounts by the desired level of privacy and security. Florencio & Herley [9] conducted a large scale study of password use through the six-month deployment of a Microsoft toolbar. They collected data from more than 250,000 users, and found that the average user had 6.5 passwords, each of which was shared across 3.9 websites. They found that the average user accessed 25 accounts over the six month period, and logged into eight accounts per day. A 2011 diary study of password use by Hayashi & Hong [11] collected detailed records of password entries over a two-week period. They found that users accessed a mean of 8.6 accounts over two weeks, and estimated that most participants had about 11 accounts in total. Although they did not study password reuse directly, all of their participants reported reusing passwords across multiple accounts. A more recent diary study [18] conducted in an organizational setting found that users authenticated 23 times a day on average, and were frustrated by the frequent disruptions to their primary tasks.

Password-composition policies also influence how users choose and manage passwords. A study of a change in password policy at Carnegie Mellon University found that the shift to a more complex policy annoyed and frustrated users, causing them to rely on new coping strategies, but also made them believe that they were more secure [17]. Another study showed that password policies do influence how users choose passwords [13]. However, users are likely to retain fragments of existing habits and passwords across changes in policy, leading to long-term reuse [21].

While several studies have investigated *what* users do to cope with passwords, there exists less investigation into *why* users behave the way they do. Wash [22] identified folk models of security threats (viruses and malware) that users use to justify ignoring security advice. A follow-up study [15] investigated how users find information about security, and found that most users depend on informal shared security stories from friends and family.

Figure 1: Participants were provided with cards showing website screenshots, to help them understand and immerse themselves in the task.

Another problem in the deployment of useful security advice is the divide between those who make security policies, those who enforce security policies, and those who follow security policies. Studies of information security in organizations have revealed a "digital divide" between policy makers, who typically do not bear the cost of security vs. users, who handle the downsides of security including lost productivity and opportunities [3]. Beautement, Sasse & Wonham [5] suggest that organizations must factor in and budget for the cost of employee compliance with security policies. They suggest that organizations need to consider costs and benefits to the organization when setting security policy.

3. STUDY

To investigate how users manage and keep track of their passwords, we conducted a series of interviews about password habits. The interviews were facilitated by the researcher, who asked the questions, recorded answers, and encouraged participants to discuss or give fuller answers. The interview was audio-recorded to allow further note-taking and analysis. We also conducted a brief self-administered demographics questionnaire that collected basic information including age, gender and occupation, and was mostly intended to give a better understanding of the interview sample. This study was approved by our ethics board.

We developed our interview questionnaire around the idea of exploring users' password management techniques. We asked a set of general questions about password habits and usage, including questions about how many passwords and accounts participants had, whether they reused passwords, whether they used password managers, and how they kept track of their passwords. The next series of questions asked about how they would behave when creating new accounts, and when changing or resetting the password on an existing account. We did not ask participants what their passwords were, and we specifically told participants that they should never reveal their passwords to us. Each interview took approximately 30 minutes, and the interviews were conducted at our university.

We chose our methodology to encourage participants to discuss thoughtfully the ways in which they approach the task of password management. We used a guided interview to focus the discussion around topics of interest to us, but we asked additional questions to probe responses and follow up on emerging topics of interest. We broke questions into a number of parts to give participants an opportunity to fully explain how and why they make their decisions, and to avoid having participants rush through their answers. We provided users with props in the form of cards with website screenshots (Figure 1) to situate themselves in the password creation and reset tasks, and to encourage them to consider their real life behaviour.

We used Grounded Theory [19] to analyze the interview data and conduct qualitative analysis of participants' responses and discussion. Grounded Theory is an analytical framework that seeks to develop an explanatory theory from a set of data. It builds a theory grounded in evidence, rather than validating an outside theory or testing generalizability. Grounded Theory defines a theory as:

> "... a set of well-developed categories (e.g., themes, concepts) that are systematically interrelated through statements of relationship to form a theoretical framework that explains some relevant social, psychological, educational, nursing, or other phenomenon." (p.22 [19])

4. RESULTS OVERVIEW

There were 27 participants interviewed for the study, and all were recruited from the university community via posters, mailing lists, and word-of-mouth. In conducting the interview, we used the constant comparative approach, where we refined the focus of the interview discussion throughout the study. At 27 participants, we reached saturation, where we were hearing little new from additional participants.

Two-thirds of participants were female. Participants' age ranged between 17 and 67, with a median age of 22. Most participants were either full- or part-time students, and came from a range of programs including the humanities, sciences, and social sciences. None of the participants were studying computer science or computer security. The other participants worked in the university community, in roles such as administrative assistants, librarians, and security guards.

In addition to the discussion and deeper responses, the interview questions also yielded a set of quantitative data about how many passwords and accounts users have, how many passwords they reuse, and the extent to which they use password managers and other tools. We present below descriptive statistics of the responses to these questions. We present these data before the Grounded Theory analysis to give context to participants' responses.

The first part of the interview investigated how many accounts and passwords users have. We wanted participants to closely reflect on their answers to these questions, so we divided questions into multiple parts. For example, in a question about number of accounts, we identified 14 account categories where participants might have accounts, and asked about each category individually. We hoped this technique would help users remember infrequently used accounts.

Participants reported their total number of accounts as being between 9 and 51 accounts, with a median of 27 accounts. The bulk of most participants' accounts consisted of email addresses, school or work accounts, and social networking accounts. They reported using a median of 11 accounts in an average week, with a range of 3 to 14 accounts.

Participants reported having between 2 and 20 unique passwords, with a median of 5 passwords. All but one of the participants in the study (26 participants, 96%) reported reusing passwords between accounts. Of the participants who reported reusing passwords, most (23 participants, 88%) reported reusing more than one password, and 19 (73%) reported reusing passwords either "always" or "frequently". Participants described different strategies for reusing passwords. Some described using the same password for all accounts, and others described linking passwords with usernames. Participants also reported using different passwords for different online contexts, such as at work or school. Several participants mentioned that they were careful not to reuse passwords on "financial" or "important" accounts (though many did not clarify what was important). Conversely, many participants also mentioned having a specific password that they reused widely on accounts of low interest, low importance, or infrequent use.

We were interested in whether participants considered context of use in their password management strategies. Participants reported entering their passwords on a range of devices including desktop computers, laptop computers, tablets, e-readers, and smartphones, but the most commonly reported context was a computer (laptop or desktop) and a smartphone. Most participants (18, 66%) said that they did not consider device constraints when choosing passwords. Participants who considered device constraints mentioned that they checked the availability of apps (to reduce the difficulty of password entry), the different security requirements of different devices, and awareness of the usability of different keyboards. All participants reported that they enter their passwords on computers that do not belong to them. Several participants mentioned that they were more careful about logging out on these computers, and about not saving passwords in the browser. One participant mentioned that they sometimes changed their passwords after entering them on computers belonging to other people.

The next set of questions addressed the coping strategies that users develop to keep track of passwords and accounts. We asked participants if they used any kind of password manager (including the browser-based managers), and 22 respondents (81%) said that they saved their passwords in some kind of password manager. All of these went on to clarify that they saved passwords in their browser or in the Apple Keychain. No one reported currently using dedicated password management software, although one participant said that he had previously used one. We also asked if participants ever clicked the "remember me" button to stay logged in to websites using cookies, and 22 participants (81%) reported clicking these boxes. Interestingly, although the same percentage of participants said they used cookies as password managers, these sets did not completely overlap.

Twenty-one participants (78%) reported writing down at least some of their passwords. Of these participants, most referred to the recorded passwords as a backup for memory, and not a resource used at every login. Participants reported different recording strategies – some recorded only part of the password, or a hint to the password, while others were more methodical about recording all of every password. Participants reported using both physical and digital media to

store passwords, but specified that the recorded passwords were easily accessible from their regular computing context.

The final part of the interview asked participants about password changes and resets of forgotten passwords. Forty percent of participants reported having ever changed passwords of their own volition, and these participants remarked that they changed passwords rarely, and only under special circumstances. Most participants evidently did not consider situations where they changed forgotten passwords, because all participants reported having done this. Most participants reported resetting forgotten passwords once per month or less, and most people said that their strategy in those cases was to change the password to something similar to existing passwords (often reusing or adapting an existing password).

5. QUALITATIVE ANALYSIS

We chose not to fully transcribe our data. Instead, we made detailed notes about responses to the interview questions. These notes included quantitative question responses, but also included additional details from participants' discussion of the topic. In places where our notes were not sufficiently detailed, we returned to the audio-recorded data for additional information. We referred to the audio-recordings to transcribe exact quotes for use in this paper.

For the qualitative analysis, we followed the grounded theory methodology of Strauss & Corbin [19]. This method involves several steps in the analysis process. First, recorded data is analyzed point-by-point and assigned descriptive codes, in the process of *open coding*. Next, these codes are compiled, and the process of *axial coding* looks for relationships among the codes. In the process of axial coding, the researcher asks questions such as *why*, *where*, *how*, and *when* in an effort to uncover structure in the data. Finally, *selective coding* integrates the results of the open and axial coding, and refines them into a theory.

5.1 Open Coding

We began the process of Grounded Theory by developing a set of descriptive open codes. We generated the open codes by examining the noted responses from the interview data. We traversed the answers to each question, looking for recurring patterns and themes in the data. Each of these themes was denoted by a code. We had a total of 66 codes.

Some of the codes emerged in relation to the question being asked in the interview. For example, we asked participants about whether they wrote their passwords down, and how they stored and referred to recorded passwords. Several codes about password recording emerged from responses to that question. However, other codes emerged over the sequence of the discussion. Participants gradually revealed more about their password creation, organization and categorization techniques as we explored how they would handle password creation, how they would choose passwords for new accounts, and how they would keep track of new accounts.

One of our password recording codes was *records passwords as backup strategy*, and we used this code when a participant indicated that although they wrote at least some of their passwords down, they did not refer to these recorded passwords on a regular basis, and instead appeared to use the recorded passwords as a backup.

In the following quote, the participant describes how she used to write her passwords down as a fallback for memory when going on vacation.

> "Not any more. I used to. [Why did you stop?] Because the only reason to write them down was if I was going on vacation for two weeks and I'd come back to work and I wouldn't remember my password [laughs]. So that was [garbled] but now I very rarely take vacation more than one week at a time and I can remember one week [laughs]." – P15

This participant describes writing her work passwords down so that she would be able to remember them after a long delay. However, she does not need this technique in everyday use. She also explains how a change in her circumstances (shorter vacations), has affected her password coping strategies (coded as *change of habit*).

Another code was *single sign-on* which we used when a participant brought up the subject of single sign-on services, such as through Facebook or Google.

> "Or you could just connect through Facebook. [It's true. Would you/do you do that?] I actually don't do that very often, no. [Why not?] Uhh, I just find there's so much junk on Facebook sometimes that I really just don't want to add to it. On Facebook, I try to only add, umm, things that are more important to me I guess." – P27

This participant shows a misunderstanding of single sign-on when she explains that she avoids signing in through Facebook because she thinks her activities will be posted to her Facebook feed. Interestingly, she does not express concerns about privacy, but rather about the relevance of information that she posts to her personal page.

5.2 Axial Coding

Following open coding, we began the process of axial coding. In axial coding, we took the codes assigned in open coding and looked for patterns, connections, and relationships between those codes. Our eventual goal was to form a model or theory that described our data. To examine the codes, we tagged each code with a post-it note, and arranged them on a table to look for connections (Figure 2).

A list of codes used in open coding is presented in Table 1, along with a brief explanation of each code. The groupings in the table are the result of the first round of axial coding where we collected codes with a similar focus. Following this, we identified an ordering, and then assembled our groupings into larger categories following this order. These categories are described in the subsequent sections.

5.2.1 Choose Your Password

At some point, every user must create their passwords and how they do this is up to them. In our interviews, participants discussed a number of strategies that they employed when choosing passwords. Some participants included personal information in their passwords. Participants mentioned including the birthdays of loved ones, phone numbers, and personal information such as hobbies in their passwords.

> "I'll try to usually think of some kind of hobby of mine, uhh, whether it would be something like hockey or a video game and a video game character, and I'll try to link it to that. Something that I usually think about quite a bit." – P24

Table 1: A selection of the 66 codes used in the open coding process. The codes are organized into related groups based on the initial step of the axial coding process.

Code Name	Description
Contextual Behaviour	
Banking as important account	Banking was a frequently distinguished important account.
Contextual behaviour for different websites	Different password behaviour on different websites.
Contextual behaviour for different environments	Different password behaviour in different places (e.g., work vs. home).
Password Categories	
Categorizes by security	Categorizes password reuse by the level of security.
Categorizes by password rules	Categorizes password reuse by the password rules on the website.
Categorizes by frequency	Categorizes password reuse by the frequency of website use.
Categorizes by semantics	Categorizes accounts by content similarity.
Passwords linked to password rules	Creates passwords to use with different password policies.
Affective passwords	Picks passwords that have emotional significance.
Personal information in passwords	Incorporates personal information into passwords.
Creating Passwords	
Algorithmic passwords	Uses some kind of algorithm to generate passwords.
Variations on a theme	Unique passwords consist of variations on a single password.
Passwords linked to usernames	Associates passwords with unique usernames
Passwords linked to times	Associates passwords to the time period in which they were created (e.g., during undergrad).
Passwords linked to website content	Links passwords to content found on the website, or reason for visiting the website.
Affective passwords	Creates passwords with emotional significance.
Personal information in passwords	Creates passwords with personal information (i.e. phone numbers, birthdays).
Named Passwords	Has a specific nickname for their most frequently used password.
Preferred Characters	When fitting password to password policies, has a set of habitually used numbers and symbols that they add.
Password Recording	
Digital Recording	Records passwords in digital media (e.g., in email, or in an excel file).
Physical Recording	Records passwords in physical media (e.g., on post-it notes, in a journal).
Records as backup strategy	Records passwords, but does not refer to them consistently.
Always records	Systematically records all of their passwords.
Records when special policies	Records passwords when the website policy prohibits resets or disables cookies.
Records clues to password	Records hints or clues about the password.
Tools	
Uses tools only in some contexts	Uses cookies or browser password managers only on some devices.
Combination of coping strategies	Uses a combination of coping strategies to remember passwords (e.g., password manager, writes passwords down, and password resets).
Unable to take advantage of coping strategies	For some reason, cannot use a certain tool or technique to remember their passwords.
Personal Validation Questions	Relies on the personal validation questions to reset forgotten passwords.
Single Sign-On	Sometimes uses single sign-on to log into different websites.
Password Rules	Creates passwords by consulting the available password policy.
Attacks on Self	
Guessing attack on self	At login, attempts to guess own passwords.
Dictionary attack on self	At login, guesses all of own reused passwords.
Password Difficulties	
Password reuse not working well	Reuses passwords, but still has problems managing or remembering.
Password reuse for memorability	Reuses passwords because unable to remember more passwords.
Security Concerns	
Privacy	Explicitly considers privacy when creating accounts.
Difficulty	Expresses the difficulty of managing and remembering passwords.
Behaviour Change	
Change of habit	Has had a major change of behaviour in how they create, manage, or remember passwords.
Hacked	Described an incident where they found the security of an account had been breached.

Figure 2: Re-arranging the codes to look for patterns in the axial coding process.

These strategies were often combined with affective strategies that included personally meaningful information in passwords. One participant told us that she had changed her passwords to a personal goal, so that the password would be easy to recall, but also so that she would be continually reminded of the goal.

> "I read an article this, this, uhh, month, that said 'whatever your goal is, make that your password' [okay] and you can still follow their rules . . . but because you're going to be entering in your password so many times a day, make it your goal, and it can be anything, you know." – P15

Another participant said that she included religious phrases such as "God is good" in her passwords, as a reminder of her beliefs and priorities.

Another strategy described was temporal. One participant told us that she changed email addresses depending on the point in her life (it appeared that she habitually switched all of her email into the address associated with her current educational institution). She had a password associated with each of her email addresses (when used as usernames). At login, she considered the time period in which she created the account, and entered the password linked with that email address and time of life.

A number of participants told us that they linked passwords closely to website content. As an example, one participant told us that if he was creating an account on an online store, he might incorporate the item he was purchasing into his password.

A few participants mentioned an algorithmic strategy for creating passwords. They systematically combined pieces of information to create passwords with a consistent format. Participants described different pieces of information that were included in their passwords. One participant said she included a piece of information associated with the website, as well as a piece of personal information in each password. Participants also had a few standard symbols, numbers, or words that they recombined for variation in their passwords.

External factors are also taken into account when choosing passwords. Several participants told us that they liked having the password policy rules displayed at create time, because they can factor them into their passwords, rather than having to create a password and modify it when it does not satisfy the policy.

5.2.2 *Reuse Your Password*

Password creation always happens with a new account, but users almost always have other accounts as well. Most of our participants reused passwords across accounts. We were interested in how they chose to reuse passwords across accounts, and how they matched passwords with accounts.

The participants in our study who reused passwords all built a personal model of reuse. Often, participants described categorizing their accounts and assigning passwords to categories. Participants described a number of different categorization strategies. Security was a common consideration that many people mentioned.

> "Like I said, it depends on what the website actually is. If it requires a weak password according to me or a strong one, I'll chose it on that basis and probably alter a letter or two." – P25

Participants assessed the security needs of the websites, and they referenced matters such as privacy, and confidentiality without clarifying those terms. Many participants explained that they treat accounts differently if they store credit card information. We were unclear on how well they assessed the security needs for non-financial personal information (such as on social networking websites).

Several participants described reuse strategies that hinged on password policies. One participant told us that he maintained a set of five passwords that fit increasingly complex password rules. If the password rules were easily viewed, he chose the appropriate password. If the password policy was not displayed, he began by trying the simplest password and only trying more complex passwords if the site rejected the simpler password.

Many participants described a semantic or thematic approach to their password reuse. They attempted to reuse passwords on accounts with similar purposes or contents. Examples included using the same password for social media websites, or across online shopping accounts. Participants also described strategies that organized passwords into less obvious semantic categories. These included using the same password on all professional accounts, or on all accounts that had low personal value.

Unexpectedly, many participants discussed frequency of use when describing password reuse. Participants mentioned having trouble remembering passwords for infrequently used accounts, but surprisingly, they often seemed to feel that infrequent use indicated a lack of need for security. Correspondingly, these same participants saw frequently used accounts as needing more protection. One way of bolstering the frequency of infrequently-used passwords was to have a single password for infrequently-used accounts, but it was unclear whether the purpose of this was to group accounts.

It was clear in the interviews that although many people had several reused passwords, there was a primary password that was reused on most accounts. Participants referenced this password in a variety of ways, but the language used indicated the importance of this password. One participant called it her "go-to password" and told us that she relied on it because she trusted the person who had chosen it for

her. Several participants referred to a certain password as being "familiar" or "easy". Many participants remarked that they had many passwords that were variations on a single password, and it appeared that these were often variations on this most used password.

> "[How many unique passwords do you have?] Eight? But it's always, like, you know, adding a one at the end when I forget. [So some of them are slight variations?] Yeah, yeah." – P9

A few participants were unable to describe any particular strategy to their reuse, although they definitely did reuse passwords. One participant told us that they "randomly" choose one of their reused passwords, and another participant said that they cycle through their reused passwords in order when creating accounts. None of these participants mentioned any reasoning for their habits.

5.2.3 Commit Your Password

After assigning a password to an account, the user must be able to keep track of this password. In our study, participants described a variety of coping strategies that they used to remember (in the active sense – store) their passwords.

Writing Passwords Down: The majority of participants told us that they wrote at least some of their passwords down. Some participants described strategies where they recorded all of their passwords, and others told us that it was a strategy that they used only in special cases. Most participants said that they wrote their passwords down to prevent forgetting them, but others wrote their passwords down as part of a larger strategy. One participant told us that she records her passwords in a spreadsheet for her husband to access in case of emergency. She later implied that she sometimes consults the spreadsheet for herself, but this is not the primary reason for keeping it.

> "[Do you ever write your passwords down?] Only maybe a couple of banking ones, they're the only ones, my banking, so if I die my husband can find them." – P5

Most participants appeared to view their password recording as a backup strategy rather than a constant resource. Interestingly too, writing passwords down can support not needing to rely on such techniques forever. One participant told us that she wrote passwords down only until she had memorized them. She also used a rehearsal strategy to help her memorize her passwords.

> "If it's new, I'll write it down for the first couple of times, but if it's new, I'll try to remember it, try to memorize it. I'll log in a bunch of times until I've memorized it." – P11

A number of participants described special cases where they would write passwords down. These special cases included assigned passwords, websites without any backup mechanisms (such as online password resets), and websites where the use of cookies is disabled. Other participants told us that they recorded hints or clues to their passwords. A few participants told us that they recorded usernames, either with or without the corresponding password. Another strategy was to write down only part of the password or some other form of password hint.

> "I email them to myself, and I have a folder with all my passwords in it [This is an email folder?], (nods) and instead of putting the actual password in there, I put something to remind me what the password was." – P13

An important consideration in the safety of recording passwords is how they are stored. Participants in our study described a variety of storage strategies for recorded passwords. Some participants wrote down their passwords on physical media, such as post-it notes, or journals. Often, they chose places near their computers to keep these passwords. One participant told us she pins a list of passwords to her bulletin board, while another described writing her passwords on a box kept on her desk. Some participants went to lengths to hide these passwords, by keeping the post-its underneath keyboards or by carrying the list of passwords with them, but others seemed unconcerned. Participants also shared a variety of strategies for storing their passwords digitally. These strategies included dedicated password documents (Excel spreadsheets, or Word documents), passwords emailed to themselves, passwords kept in online notebooks (such as Evernote), and passwords stored in desktop widgets.

The accessibility of recorded passwords was a key issue. For participants who chose to store their passwords digitally, they often mentioned concerns about having the password list when it was needed. A few participants mentioned using services like Dropbox to keep their password lists in the cloud, and participants who emailed passwords to themselves appeared to have chosen this technique for accessibility. One participant emphasized the need for accessibility when he described using services that synced across devices to store passwords.

> "[Do you ever write your passwords down?] On my phone, sometimes. ... Usually you would either keep it in, like, Google Keep, or iCloud, or messenger, or Evernote. In my case, Evernote. I use a lot of Evernote, so.... Anything that is really sync-able to multiple devices that way it is easier for me to store info." – P16

Password Managers: Almost all of our participants reported using the password managers built into web browsers. A few participants told us that they only used these tools in specific contexts: two participants told us that they saved their passwords in the browser at work, but not at home, and many participants clarified that they only saved passwords on their own computers. A number of participants also clarified that they only saved passwords for certain accounts in the password manager. Most commonly, participants said that they did not save the passwords for banking websites, but a few people specified that they did not save passwords on any websites that required credit card details.

One participant in our study told us that he used to use a dedicated password manager, but had stopped using it. He was vague on the details of the password manager, and could not remember its name, but was able to tell us that he had stopped using it because it was inconvenient to have to copy passwords out of the password manager.

> "I've used, umm, I can't recall the name, but I've used one before in the past. [What made you start or stop using it?] It was just inconvenient

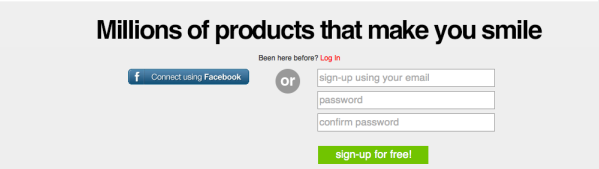


Figure 3: A detail of one of the websites used as a prop in our interviews.

> because it encrypted it every time and I'd have to decrypt it. [Oh, okay] It was the same issue, it's like encrypting and decrypting constantly. It's just more convenient to have it. [And was that kind of like the time that it took to...] Yeah, but you had to basically sorta copy it, paste it into a search box. It wasn't the time issue, it was just the hassle if I wanted to do it with multiple passwords." – P25

This participant refers to the time and hassle of encryption and decryption, but we speculate that he is referring to the process of hashing a master password.

Many websites offer the user the opportunity to remain logged into a page via a cookie saved in the browser. This choice is often presented at login, via a checkbox that says "Remember me", "Stay signed in", or something similar. Although this mechanism is not one that saves passwords, it accomplishes the same result by removing the need for the user to enter a password. In our interviews, we asked about cookies directly after asking about browser-based password managers, and it was clear that a number of participants did not understand the difference between the mechanisms.

> "[You know how sometimes you'll go to log in, and there will be a box for your username, a box for your password, and then there'll be a little box to tick that says "Remember me"? Do you ever tick that box?] Yeah. Because that's the same as saying 'save your password', right?" – P16

Some participants appeared to treat the mechanisms identically (typically, relying on both), but others told us that they used one or the other without giving much justification for their habits.

Other Tools: Over the course of the interviews, participants mentioned a few other tools and techniques that they used to keep track of passwords. These tools included a Smartwallet app, and single sign-on. Although a single sign-on option ("Connect via Facebook" – see Figure 3) was prominently displayed on the password creation prop card, only two participants commented on it. One participant told us she would use the option when it was available because she had a hard time remembering even her reused passwords. However, another participant (quoted in Section 5.1) said she would not use it, because she did not want any extra information cluttering her Facebook page. It is difficult to know why other participants did not mention any kind of single sign-on, since the cue was equally visible to all participants. Our interview script did not prompt them to specifically look at the Facebook button, but they were told to imagine they were on that page. Possibly, this indicates that most users do not understand how single sign-on services can be used as an alternative to reusing passwords.

In summary, it was clear that most participants used a variety of tools and strategies to help them cope with their passwords. Participants saved passwords in browser-based password managers to handle the case when they were using their own computers, but needed backup mechanisms for times when using other computers.

5.2.4 *Forget Your Password*

After a user has memorized their password, there is always the chance they may forget it. Users have many passwords, and it is clear that handling forgotten passwords is a large part of the password management task.

Several participants described situations where they could not remember their passwords at login, and told us their first action would be to try and guess their password. Some participants described a kind of targeted dictionary attack on themselves, where they would guess all of their reused passwords.

> "Sometimes, I do forget, but I try everything else [all of the passwords]" – P17

Other participants described guessing strategies where they attempted to recreate their motivation for being on the site (for example, the item they were buying when they created a shopping account), or the password they would have been likely to pick in the time period they created the account. Still others said they would try to recreate algorithms for password creation, or look at the password policy to make a better guess at their own password.

The other fallback strategy that participants described was the password reset mechanism. In this, we include both personal verification questions and email resets. Almost all participants told us that they have reset forgotten passwords, and it appears that many users do this on a regular basis. Some users seem fine with this as a strategy, but others raised objections. One participant told us that she had begun writing her passwords down when she realized she was resetting her passwords too often. Another participant remarked that she had not considered how often she reset her passwords until the interview, but that it was a major part of her password coping strategy:

> "It's funny, I never really thought about it, but I guess I do that a fair amount." – P19

5.2.5 *Live with Your Passwords*

Passwords and accounts can last a long time: once passwords have been created and linked to accounts, all users must begin the long process of living with and coping with their passwords.

A number of participants in our study commented on the difficulty of managing and remembering passwords. Participants displayed different attitudes toward this difficulty. One participant referred to passwords as agonizing.

> "Then what do I do? Ohhh, my gawd. Then I agonize for a few minutes." –P13

Another participant seemed resigned to reusing passwords.

> "[Do you ever reuse passwords?] Oh yeah! (laughs)" – P9

Participants also referred to fears of doing the wrong thing, and uncertainty about the outcomes of password decisions. Some participants explicitly referred to privacy, but only a few referred to private information that was not financial.

As time progresses, a user may change their behaviour in some way. In our study, a number of participants described changes in behaviour that had occurred for a variety of reasons. Some users described stopping or starting using tools. One participant told us that she no longer saved passwords in the browser because she had heard that this behaviour could be dangerous.

> "I used to and then I took one of these studies and they helped me with understanding why that was not a good idea. [Did you at that point go and erase the ones that were already in there?] Yes." – P13

Another participant told us that he used to use a password manager, but had stopped using it because of the time it took to copy and paste passwords when using the password manager. He describes the hassle of using the password manager in a quote in Section 5.2.3.

One reason to change an existing password is in the case of a security breach. A number of participants in our study described situations where they had changed account passwords after suspecting that an account was not secure. Examples of breached accounts included email and PayPal accounts. As participants described their security breaches, it emerged that they were often unsure about whether they had been attacked, and sometimes had to made decisions without really knowing what had happened.

> "At least, I think that I've been hacked. [What kind of clues, what would be a kind of signal to you? Has it ever happened to you?] It has, because my friends told me they got these really strange emails, from my email, supposedly sent from me, that were obviously ads for something or other and they were like 'hey, this doesn't sound like you'." – P19

A few participants brought up changing their passwords in the case of more minor suspected security breaches. One participant told us that they changed their passwords when they thought a friend might have seen their list of passwords, and another participant said they had changed their Facebook password when they thought they might have left the account logged in on a friend's computer.

5.3 Selective Coding

The last coding step in Grounded Theory is selective coding, where the researchers attempt to identify a unifying *core code* that describes the underlying phenomenon in the observed and interpreted behaviour.

As we analyzed the data, a central theme about rationing and budgeting began to emerge. In all phases, our participants described ways in which they stretched thin resources: memorization, attention, creativity, and security knowledge. Similar to the way in which we ration and conserve time, energy, food, and money, participants were handling password management by devoting appropriate resources to accounts of great importance, and then devoting less energy to other accounts, and generalizing their approach to similar accounts to save effort. In their work on organizational security, Beautement et al. [5] suggest that organizations need to budget for the costs (both time and money) of organizational compliance. Our suggestion about rationing differs in that we identify that individual users are budgeting their own time and effort. We are not suggesting this arises because of a lack of willingness to comply, but rather from a paucity of cognitive resources.

In the following sections, we systematically examine how rationing plays a role in each of the themes of axial coding.

5.3.1 Choose Your Password

When choosing their passwords, participants rationed their efforts in a variety of ways. For participants with formulaic or algorithmic strategies, part of their investment was in memorizing their personalized strategy. By remembering that their strategy is to include a word related to the website, they reduced the amount of effort that it takes to choose a password on a new website. Participants who closely associated their passwords with a username were engaging in a similar strategy. Remembering usernames can be difficult, but if you have a consistent strategy to associate a password with a username, effort can be more effectively rationed to each account.

The interviews asked participants what they would do when creating an account if their password was rejected on the grounds of insufficient complexity (for example, lacking a symbol). Most participants reported that their strategy in this situation was to append a symbol to the password. Most participants referenced "their" symbol, and told us that they had a habitual symbol that they used in this situation. This coping strategy implies a way of rationing effort across situations that cannot be predicted. If participants knew their password would need a symbol, they would have begun with a symbol. But since they are unable to see the password policy, they have developed sensible coping strategies that conserve memory and effort in these situations.

Many participants told us that they reused pieces of passwords in a variety of ways. Many participants referenced appending different endings onto the same widely reused passwords. Participants also discussed using this strategy as part of their password changes; one participant told us that after he realized a friend might have got access to his list of recorded passwords, he had changed his passwords, but had simply added characters to the existing passwords.

> "Like I said, I have them all stored on a text file, right. And once, a friend of mine or an acquaintance borrowed that USB drive, and I felt that he would have access to everything so I went and changed everything. [Okay] But all I did was add like a letter or two." – P25

5.3.2 Reuse Your Password

One main way in which users ration their efforts is in not choosing a new password for every account. Reusing passwords allows users to conserve energy across their large number of accounts.

As participants discussed how they would create a password for a new account on an online shopping website, many digressed into a discussion about accounts that do not matter to them. One participant told us that she had a password that she used on accounts where she would not care if she was hacked. Another participant referenced having a pass-

word that she would not mind sharing with others. Whether or not these participants actually would not care if others had access to their account, their behaviour shows that they are rationing the effort they put into these accounts by the amount of effort they merit.

An important aspect of rationing is that those who deserve more should get more, and we found evidence that users were applying this principle in their password management strategies. Many participants referenced special habits for their online banking: participants told us that they did not reuse their banking passwords, that they would not log into their bank on a shared computer, and that they would not enable cookies or save their banking login in the password manager. All of these behaviours indicate that users are willing to ration more effort into accounts that they perceive as needing it.

Participants referenced using different behaviour in different contexts. Two participants referenced a different set of habits for their work passwords, and both implied that they were less careful about security for these passwords. It was difficult to know if this was because the information was seen as less personal, or because they felt that the physical environment was more secure, or if it was simply because they used those accounts more frequently, but they were assessing information about the context of use and rationing effort differently to it.

5.3.3 Commit Your Password

Memorizing passwords is one of the most difficult parts of the password management task for users. Passwords must be maintained over long periods of time, with sporadic and unforeseeable usage patterns. Using tools such as password managers and techniques to write passwords down is how users allot effort into the unknown needs of the future.

Availability and accessibility are key issues for password managers. Managers are typically only available on personal computers. Dedicated password manager software sometimes have associated smartphone apps that allow users to take their passwords away, but the browser-based managers are largely only tied to one computer. This means that users must ration the effort they put into making sure their passwords are available to them when they need them.

Participants in our study described using a combination of techniques to keep track of their passwords. Some participants were heavily invested in one strategy, but most participants appeared to know a few of their passwords, to have some of them written down, and to have some of them stored in a password manager. This strategy seems to ration effort across time and place – when at home, the password manager saves the passwords for almost all of their accounts, or they might have easy access to the recorded passwords. When elsewhere, they cope by remembering their passwords, or by carrying some of the passwords with them. It appeared that some participants were sacrificing convenience for security in these situations, which indicated another aspect of rationing in their coping strategies.

5.3.4 Forget Your Password

All of the participants in our study told us that they have reset passwords when they are forgotten. Participants clearly regarded this as being separate from a change of password, which seems to indicate that forgotten passwords are seen as part of the landscape of password management.

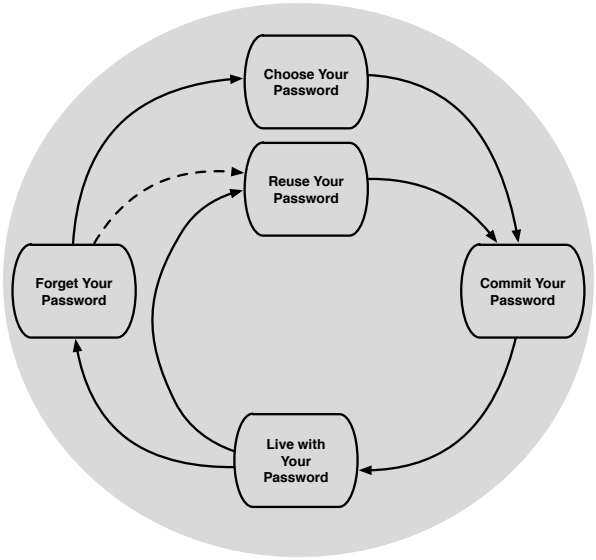


Figure 4: The password life cycle.

Users expect to have to handle a loss of memory, or a failure of a coping strategy. It appears that the password reset mechanism allows users some flexibility in their rationing strategy. In particular, many participants appeared to rely on email resets. In situations where a password reset was not available, participants described rationing extra effort to the situation, often to ensure that the password was recorded.

In the situation where they were resetting a password, many participants told us that they would reuse an existing password, or change the password to what they thought it had been or what it should have been. It appeared that their coping strategies were equipped to handle these situations without any particular sense of frustration or loss.

5.3.5 Live with Your Password

Throughout the interviews, we heard a number of remarks on the difficulty of password management. Although participants are resigned to the realities of passwords, they still present difficulties. Finding and implementing coping strategies for the difficulties of passwords involved effort and unpleasantness. Similarly, rationing itself is a difficult task! The decisions about how to allot time, energy, and effort are not always obvious to users. Users referenced lacking information that would have made it easier to cope: unseen password policies, misunderstood security requirements, and invisible security breaches.

5.4 The Password Life Cycle

In the final step of Grounded Theory, we look back at the identified codes, patterns, and relationships in order to form them into a theory.

Our theory is that there is a password life cycle – a progression of stages through which every password passes. Passwords are created, assigned to an account, then recorded or memorized, lived with, and then potentially forgotten. Old passwords are then reused or adapted in the creation of new passwords, and the cycle continues.

Figure 4 illustrates the stages of the password cycle. The cycle begins when the user needs to create a password for a new account. Theoretically, a user might begin with no passwords at all, and have to fabricate one from scratch, but they may also have existing strategies and password phrases that they will integrate into a new password. This password must next be committed, either memorized or recorded, so that it can be later used for login. Assuming the commitment process is successful, the user then lives with their password. They login and access their accounts successfully. If they successfully remember their password, and it is appropriate for reuse, they can then reuse that password. If the password must be changed (because it is forgotten, because someone else has learned it, or because of enforced password change policies), they must return to password creation.

Rationing is present at every step of the password life cycle. Users ration effort at creating new passwords, they reuse passwords to put more protection on the most valued accounts, they reduce the effort of memorization by saving passwords in managers or by writing them down, and they strategically budget the attention they pay to passwords on existing accounts. Users save resources from inconsequential accounts so that they can devote them to to more important accounts. Allotting time, attention, and energy to different accounts forms the backbone of users' coping strategies. As with other forms of rationing, users scrimp on effort for some accounts to save it for others.

Rationing contributes to the cycle of password reuse. As effort is reduced from some accounts, it is saved for new ones. Reused passwords are handed down from existing accounts, saving the user the time and energy of creating and memorizing a new password.

6. DISCUSSION

Our theory suggests a number of ways in which the design of security products could better support users.

6.1 Writing Passwords Down

While writing passwords down is an intuitive and reasonable way of handling security, users need helpful guidance on the right way to store these passwords. Writing passwords down is conventionally understood to be insecure, but many security experts actually advocate writing passwords down [6, 16] if they can be kept in a physically secure location. Many users do write their passwords down, but the caveat about storage is poorly understood by users. In our study, participants reported keeping their password lists in their email, in Dropbox, on their cell phones, or saved on their computer desktops. Recording passwords is a sensible way of conserving effort, and users should be encouraged to make the small changes that could make this habit safer.

To address the storage problem and guide users to safer password storage, a plausible solution might be the development of a service specifically to securely store passwords. Password storage notebooks do exist (*e.g.* The Personal Internet Address & Password Log Book [1]), but there is no equivalent online service. In the absence of a trustworthy electronic service, it seems possible to better emphasize the notion of physical security of stored passwords to users, and suggest secure and sensible places to keep lists of passwords.

6.2 Password Cues

Another finding of our study is that participants often had trouble matching passwords to usernames or to websites. If users were able to better match their passwords to websites, they would not have to resort to strategies where they reveal multiple passwords to potential attackers by trying all of them at login. We suggest that image cues could help users better associate passwords with accounts. Websites could be designed to associate these cues with usernames and present them at password creation and at every login. Since the cues would have no relation to the password, it would not be necessary for them to be secret. This could be similar to the image-based anti-phishing mechanisms found on some financial websites (for example, Bank of America's SiteKey [4]). The explicit nature of these cues would help users associate passwords with websites. Although some users currently try to create a cued match from password to website by including website-related information in their passwords, these strategies can place an additional burden on the user because they must then remember the cueing strategy in addition to the cued password. Reducing this burden would allow users to transfer effort into other aspects of the password task.

6.3 Password Managers

We were surprised to find that none of our participants used a dedicated password manager, and surprised even that not everyone was using the browser-based password managers. As long as they are well-designed, and do not store passwords in the clear, password managers seem to offer one of the best solutions for password management: comprehensive, convenient, and safe. However, most of our participants appeared unaware of prominent password managers, and some participants expressed distrust in this software. We suggest that the better integration of password managers into operating systems and browsers would help with both visibility and trust. A few participants in our study did mention having passwords saved in Apple's iCloud Keychain, although it wasn't clear that any of the users were taking advantage of the password creation mechanism, or the cross-device capabilities.

We speculate that password manager software could be improved to integrate password cues and to facilitate reuse. We know that not all websites will want to integrate these kinds of features into their existing mechanisms, and the advantage of a password manager is that it is controlled by the user, and does not require changes to existing websites. A well-integrated password manager would let users ration effort into a mechanism that genuinely kept them safer.

6.4 Single Sign-On

Although the use of single sign-on would address many of users' password problems, the participants in our study appeared either unaware of or ill-informed about single sign-on. Although a single sign-on "button" was equally visible to all participants in our interviews (on the website screenshot), only two commented on it at all. One said she used it sometimes because she had trouble remembering even her reused passwords, but another participant explained that she would not use it because she did not want information cluttering up her Facebook page.

Her explanation showed a strong misconception about how single sign-on works. Instead of understanding that Face-

book's role in single sign-on is to verify your identity, she thought that she was signing into Facebook and using the online store as a part of Facebook. This indicates a need for independent single sign-on providers, who do not have a stake in personal information. The biggest current providers of single sign-on are Google and Facebook, both of whom have other reasons to be interested in browsing and usage patterns. Although users do not completely understand how single sign-on works, they do not want their activities to be visible to uninvolved parties. A better option for single sign-on providers would be independent entities who only verified authentication attempts, similar to the certificate authorities who currently issue SSL certificates.

Addressing the issues with existing single sign-on services would allow users to take advantage of these services, and allow them to better ration and conserve their password efforts. Single sign-on gives the user the advantages of reusing passwords without the risk, and allows them to harness existing strategies while remaining more secure.

6.5 Extra Information

When discussing information that participants look for when creating passwords, a few participants mentioned password strength meters and a number of participants brought up the subject of password policies. Participants wanted to know password rules before picking a password so that they didn't have to waste effort creating the "wrong" password, and wanted the additional guidance of a password strength meter. Providing strength meters and making password rules available before they are broken are minimal efforts that could simplify users' password experiences.

Throughout the discussion, but particularly when they were discussing security breaches, participants referenced a lack of information about their passwords and accounts. When discussing suspected hacks, they expressed uncertainty about whether they had been hacked, and an absence of information to turn to. This lack of feedback is an inherent characteristic of security [23], but we still think that more information could be made available to users. Most websites log information about sign-ins and actions, and this information could be made available to the user. Obviously, this information could not assist in attacks where the attacker gains complete control of the account, but in many cases, participants still had access to their accounts (and were able to change their passwords) and could have used a resource to help them find additional information about account usage.

Having to search for clues about malicious usage is yet another security task that consumes users' time and resources. Making this information readily available would allow users to sensibly ration and conserve their efforts when handling compromised accounts or passwords.

6.6 Threat models

One of the emergent themes during the interviews was confusion about threat models and the nature of the threat. Although worried about security, participants seemed unclear about the type of threats that concerned them. They did not differentiate between targeted personal attacks, anonymous large-scale password hacks, and the loss of private data, although they referenced all three during the discussions. Correspondingly, participants did not seem to appreciate that the defences for different attacks might vary based on the nature of the account in question. This lack of understanding has an impact on how users ration their password efforts. If they are confused about the type of threat, they may misdirect their efforts, leaving valued accounts unprotected and over-protecting less vulnerable accounts.

Understanding of threat models informs how users categorize their accounts for reuse. In order to reuse passwords safely, users must be able to better assess their security needs on websites. If users are going to reuse passwords, either by themselves or with support from a password manager, it is important that they understand the consequences. For example, it is probably unwise for users to reuse passwords from high-importance accounts on low-importance accounts.

Assessing the threat models is important, but users are also trading off their time. Herley [12] points out that if users were to follow all given advice, the security benefits would be swamped by the time spent following the advice. He proposes instead an economic model where both threats and benefits are assessed. In order to consider both threats and benefits, users need to be able to reason about the severity and likelihood of threats, and to consider carefully benefits such as uninterrupted routines and time that can be devoted to primary tasks.

7. CONCLUSION

In this paper, we presented our study on how users cope with the difficulties of living with passwords. We conducted interviews, and performed analysis using the Grounded Theory methodology. We found that users have complex coping strategies that combine a variety of tactics. They ration effort to devote resources to accounts they feel of greater importance and minimize their effort for accounts of lesser importance. Over time, this leads to a life cycle of password usage whereby passwords are developed, reused, and adapted.

We suggest that our findings indicate new opportunities for better supporting users, and we describe some possibilities. For example, password managers might be designed to facilitate safe reuse. A proactive alternative to strength meters could help users pick appropriate passwords for each account. Accounts could help users by providing cues and password rules to help users link passwords with accounts.

The work described here is theory-building, rather than theory-validating. A large scale survey could investigate the generalizability of these findings, and see how they are supported in the larger population. Future work in this area might include the development of a survey instrument to investigate how passwords pass through the life cycle. Additionally, the information collected in this survey is self-reported. A study investigating the match between users' reported security behaviours and their real life habits could shed light on users' ideas of what they are supposed to do, as well as provide more concrete detail for design.

Our contribution in this paper is the identification of important patterns underlying user coping strategies. Users are not stubbornly refusing to follow password advice, they are instead carefully managing their resources to cope with impossible demands. Their solutions are often flawed, but they deserve consideration and may indicate better strategies for security. In choosing where to build roads, it may be best to pave the paths that users already walk.

8. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers and our shepherd for their help in improving the paper. We would also like to thank the Government of Canada for support from NSERC ISSNet and the GRAND NCE, as well as support from an NSERC Canada Graduate Scholarship, and an NSERC Discovery Grant.

9. REFERENCES

[1] The Personal Internet Address & Password Log Book (Organizer): Peter Pauper Press. `http://www.amazon.ca/Personal-Internet-Address-Password-Organizer/dp/1441303251`.

[2] A. Adams and M. A. Sasse. Users Are Not The Enemy. *Communications of the ACM*, 42(12):40–46, Dec. 1999.

[3] E. Albrechtsen and J. Hovden. The information security digital divide between information security managers and users. *Computers & Security*, 28(6):476–490, Sept. 2009.

[4] Bank of America. SiteKey at Bank of America. `www.bankofamerica.com/privacy/sitekey`.

[5] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: Managing security behaviour in organisations. In *Proceedings of the 2008 Workshop on New Security Paradigms*, NSPW '08, pages 47–58. ACM, 2008.

[6] W. Cheswick. Rethinking Passwords. *Queue*, 10(12), Dec. 2012.

[7] S. Chiasson, P. Van Oorschot, and R. Biddle. A Usability Study and Critique of Two Password Managers. *15th USENIX Security Symposium*, pages 1–16, 2006.

[8] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. F. Wang. The Tangled Web of Password Reuse. In *NDSS 2014*, 2014.

[9] D. Florencio and C. Herley. A Large-Scale Study of Web Password Habits. In *International World Wide Web Conference Committee (IW3C2)*, May 2007.

[10] S. Gaw and E. W. Felten. Password Management Strategies for Online Accounts. In *SOUPS '06: Proceedings of the Second Symposium on Usable Privacy and Security*. ACM, July 2006.

[11] E. Hayashi and J. Hong. A diary study of password usage in daily life. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Request Permissions, May 2011.

[12] C. Herley. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *NSPW '09: Proceedings of the 2009 Workshop on New Security Paradigms*. ACM, Sept. 2009.

[13] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of Passwords and People: Measuring the Effect of Password-Composition Policies. In *Proceedings of the 29th Conference on Human Factors in Computing Systems (CHI)*, New York, USA, 2011.

[14] LastPass. The Last Password Your Have to Remember, 2014. `www.lastpass.com`.

[15] E. Rader, R. Wash, and B. Brooks. Stories as Informal Lessons about Security. In *SOUPS '12: Proceedings of the Eighth Symposium on Usable Privacy and Security*. ACM, July 2012.

[16] B. Schneier. Choosing a Secure Password, Feb. 2014. `http://boingboing.net/2014/02/25/choosing-a-secure-password.html`.

[17] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. M. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering Stronger Password Requirements: User Attitudes and Behaviors. In *SOUPS '10: Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM, June 2010.

[18] M. Steves, D. Chisnell, A. Sasse, K. Krol, M. Theofanos, and H. Wald. Report: Authentication Diary Study. Technical report, National Institute of Standards and Technology, Information Technology Laboratory, Gaithersburg, MD, Feb. 2014.

[19] A. Strauss and J. Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Thousand Oaks, California, 2nd edition, 1998.

[20] S.-T. Sun, E. Pospisil, I. Muslukhov, N. Dindar, K. Hawkey, and K. Beznosov. What Makes Users Refuse Web Single Sign-On?: An Empirical Investigation of OpenID. In *SOUPS '11: Proceedings of the 7th Symposium on Usable Privacy and Security*, USA, 2011. ACM.

[21] E. von Zezschwitz, A. Luca, and H. Hussmann. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In *Human-Computer Interactionâ INTERACT 2013*, volume 8119 of *Lecture Notes in Computer Science*, pages 460–467. Springer Berlin Heidelberg, 2013.

[22] R. Wash. Folk Models of Home Computer Security. In *SOUPS '10: Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM, July 2010.

[23] A. Whitten and J. D. Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *USENIX Security Symposium*, pages 169–183. Carnegie Mellon University, Aug. 1999.

Crowdsourcing Attacks on Biometric Systems

Saurabh Panjwani*
Independent Consultant, India
saurabh.panjwani@gmail.com

Achintya Prakash
University of Michigan, USA
achintya@umich.edu

ABSTRACT

We introduce a new approach for attacking and analyzing biometric-based authentication systems, which involves crowdsourcing the search for potential impostors to the system. Our focus is on voice-based authentication, or speaker verification (SV), and we propose a generic method to use crowdsourcing for identifying candidate "mimics" for speakers in a given target population. We then conduct a preliminary analysis of this method with respect to a well-known text-independent SV scheme (the GMM-UBM scheme) using Mechanical Turk as the crowdsourcing platform.

Our analysis shows that the new attack method can identify mimics for target speakers with high impersonation success rates: from a pool of 176 candidates, we identified six with an overall false acceptance rate of 44%, which is higher than what has been reported for professional mimics in prior voice-mimicry experiments. This demonstrates that naïve, untrained users have the potential to carry out impersonation attacks against voice-based systems, although good imitators are rare to find. (We also implement our method with a crowd of amateur mimicry artists and obtain similar results for them.) Match scores for our best mimics were found to be lower than those for automated attacks but, given the relative difficulty of detecting mimicry attacks visá-vis automated ones, our method presents a potent threat to real systems. We discuss implications of our results for the security analysis of SV systems (and of biometric systems, in general) and highlight benefits and challenges associated with the use of crowdsourcing in such analysis.

1. INTRODUCTION

Biometric-based authentication is one of the most compelling alternatives to passwords for enabling access control in computing systems and, more generally, for identity management in systems. Even with some of the deployment difficulties associated with biometrics as compared with passwords, their usage in mainstream applications like banking and border security control is growing and new forms of biometrics are being continually experimented with for user authentication tasks [4].

Amongst many other reported advantages of biometrics, it is often claimed that they have an upper hand over passwords in their resilience to being faked or spoofed by ordinary human beings, even those who are acquainted with attack victims. This is also cited as a primary reason for preferring them over passwords or tokens in real deployments [8, 3, 25]. However, rigorous research on such claims is still lacking and even with a rich and mature literature on biometric-based authentication, there is no convincing answer to this simple question: for an authentication system $\mathcal{A}$ trained on biometric features of a set of users S, drawn from a large universe U, is it likely that users in S can be impersonated by those in U? In particular, is it likely that the biometric features of some user $u \in S$ are "similar enough" to those of another user $u' \in U$ for u' to be able to impersonate u to $\mathcal{A}$? This question, though generally relevant to biometric-based systems, is particularly interesting for behavioral biometrics, which define identification features over user actions (e.g., speaking or writing): such biometric forms can be "copied" with conscious human effort and differences in inherent characteristics could potentially be compensated for by such imitation.

In this paper, we consider the potential of imitation as a means to thwart biometric-based authentication systems with a primary focus on voice-based authentication or speaker verification. Speaker verification (SV) systems are gaining prominence in the real world because of the widespread use of mobile devices (numerous known deployments by banks and mobile operators; see Sect. 2) but security analysis of such systems has been limited to the use of automated tools and techniques (like voice conversion, record-and-replay) as attack vectors. In contrast, the ability of humans to imitate other humans' voices for the purpose of impersonation is less understood and generally assumed to be difficult in practice [11, 18]. Reflecting this contrast, defenses against automated attack techniques in SV schemes have become stronger with time but those against imitation attacks are still unknown.

We make two key contributions in this paper. First, we present a new method to execute imitation attacks on SV systems involving a large number of untrained users as imitators; and second, we analyze the effectiveness of this method with respect to a well-known and commonly-used SV scheme based on Gaussian Mixture Models (GMMs). The method

*Part of this work was done when the author was employed with Alcatel-Lucent Bell Labs.

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

we propose is simple and generic and it essentially involves the use of crowdsourcing to search for and identify candidate mimics for users in a given target set S. It is generic in that it does not assume a specific implementation of the SV system, except that it allows black-box access to the attacker. (Black-box access is used to identify "close matches" between candidate mimics and the targets.) It is efficient in that it uses mobile phones and crowdsourcing to quickly collect speech samples from geographically-dispersed individuals and to select candidate mimics from a large set of untrained users. We do not know of any prior work which uses crowdsourcing for biometric security analysis, voice-based or otherwise, or for analyzing authentication schemes in general. The very idea of identifying candidate impersonators from a large pool of untrained users (as opposed to hand-picking them from an expert population) does not seem to have been rigorously experimented with prior to this paper.

Our analysis of the technique with respect to a GMM-based SV system yields three key outcomes. Our first learning is that mimicry is a rare skill and that the average user of Web-based crowdsourcing platforms does not have the ability to pick the right speaker to mimic from a target set *and* to mimic that speaker well, even when provided high monetary incentives. This is somewhat expected and is also aligned with prior work which argues that professional mimicry artists exhibit greater flexibility to modify their voices than amateurs (within the realm of mimicking celebrity voices) [1, 26]. What is more surprising is the second outcome, which is that the crowdsourcing technique does identify *some* users with the ability to impersonate target speakers to the system and to do so with high consistency across authentication attempts (from a pool of 176 candidates, six achieved an overall false acceptance rate of 44%). In most cases, the imitators require help in identifying the right (closely-matching) target speaker to mimic and we found only one user who was able to self-identify a target speaker successfully. We also ran parallel experiments with a crowd of amateur mimicry artists and obtained similar success rates there, although motivating these users to participate in the experiments proved harder. Our results significantly improve upon findings from prior studies [10, 11, 15, 26] and through a careful imitator selection strategy, we are able to demonstrate better impersonation success than what has been found in these studies.

Finally, we find that even the best imitators identified by our technique fare poorer than automated attack techniques in terms of attack success rates and are unable to match the mean self-scores of target speakers in impersonation attempts. While this may appear like a negative finding, it is important to view it in the light of the fact that automated attacks are becoming easier to defend against (via different forms of *liveness* detection measures) but defenses against imitation attacks are not known in the literature. The impersonation success rates we demonstrate for our crowd-based imitators are sufficient to mount online attacks on real voice-biometric systems and current defenses for automated attacks seem insufficient to prevent them. Furthermore, given the improvement our technique offers over prior work on voice mimicry, such attacks present a potent threat to SV systems and one that future systems must suitably address. We discuss implications of our results for the design of future biometric systems (voice and otherwise) and how crowdsourcing-based analysis can assist in this process.

Before we proceed with the details, we make one important high-level remark regarding the paper. Our attack implementation should be viewed as a "proof-of-concept" of mounting crowdsourced attacks on voice-biometric systems and our work is a preliminary study of the viability of such attacks. Our main goal is to investigate whether crowdsourcing platforms with naïve, untrained users *can* be used to mount imitation attacks on SV systems and how to set up the right candidate filters to enable this effectively. The scale at which such attacks might occur on a real system cannot be deduced from our results alone. We use Amazon's Mechanical Turk to implement our proof of concept (which suffices to show attack viability) but such a platform is unlikely to be the vehicle for a real attack due to the associated legal implications and sampling difficulties in attack implementation (see Sect. 5). Further studies are needed to understand how such attacks could be implemented in practice or how the attack method could be used to analyze the security of real, large-scale systems.

The rest of the paper is organized as follows. In Sect. 2, we present some background and related work on biometric security, with a specific focus on security of voice-based biometric systems. In Sect. 3, we describe our attack technique and in the following section, we describe the experimental setup we used to implement and evaluate the technique. Section 5 presents our experimental results and the paper concludes in Sect. 6.

2. BACKGROUND AND RELATED WORK

Biometrics broadly fall into two categories—physiological, which are based on physical characterisics of an individual (e.g., fingerprints, facial features) and behavioral, which are based on behavioral traits and actions (e.g., speech, typing patterns and handwritten signatures). Speech has a unique place in this categorization in that it combines elements of both physical (vocal tract structure) and behavioral (speaking style) aspects of an individual, both of which are generally regarded to have differentiating elements across humans [13].

Biometric-based authentication systems of all types have a common structure: there is a *training* component, wherein each user submits her identity u and a set of biometric samples $\gamma_1, \ldots, \gamma_k$ to the system and the system uses these samples to prepare a "model" for u; and a *testing* component, wherein each user submits a fresh sample γ', along with her identity u, and the system checks for a "match" between γ' and the model that it prepared for u. A successful match implies successful authentication to the system. Matching is a binary classification problem—a user either classifies as u or classifies as "not u". This is different from biometric-based *identification* wherein user labels are not provided during testing and the classification task is n-ary (which of the users $u_1, \ldots, u_n$ is the closest match to γ'?). Much of the work on biometric-based authentication is around defining the right approach for modelling and matching users, which differs significantly across biometric forms.

2.1 Security of Biometric-Based Authentication

The fuzzy nature of biomterics (γ' may differ across tests even for the same user) presents new security challenges for the system designer: an adversary need not compute an exact biometric sample of u in order to impersonate as u to

the system; an "approximate" sample suffices. The system could be tuned to limit the acceptable level of approximation but this is also constrained by the fact that strict limits inconvenience real users, especially if the underlying biometric suffers from high variability across time and context (what is often referred to as *session variability*). The challenge is to come up with suitable matching thresholds which enable the right users to authenticate often enough but which cause all adversarial ways to create approximate samples to fail.

Broadly, there are two approaches to security analysis that have been considered in the literature. One involves the consideration of *automated attacks*, which use computing machinary to "create" fake biometric samples that can impersonate users to the target system. The classical automated attack is record-and-replay—digitally record samples from a user u and replay them to the system to authenticate as u. Record-and-replay is the Achille's heel of biometric-based authentication, particularly so for physiological biometrics [16, 19] which have limited scope of system-imposed dynamic variations. To defend against them, system designers normally introduce an element of freshness in the biometric capture process (e.g., for voice, have the user speak a different phrase for every authentication attempt). In the recent past, newer forms of automated attacks, like generative [2] and conversion [6] attacks, have emerged which try to defeat freshness impositions in systems by learning to generate new samples for a user u based on past samples of u and auxiliary data.

As automated attacks have grown in complexity, so have the defenses against them. Most real-world biometric systems today implement some form of *liveness* detection measures [22], which are automated ways to detect whether biometric samples provided during authentication originate directly from a human (are "live") or not. For fingerprint-based authentication, a common measure is to detect pulsation or temperature gradients in the biometric-providing object. For voice, measures range from challenge-response to the use of multi-modal techniques (e.g., capture lip-movement during authentication [7]). An emerging trend in voice-based authentication is the use of *human-mediated* liveness detection: in applications where the user is required to converse with a trusted human agent and the authentication process is incidental (e.g., phone banking), delegate the task of detecting liveness to the agent, and have the machine focus on matching[1]. Since human listeners are usually better with distinguishing machine-generated speech from human speech (and since automated techniques are not known to generate "natural-sounding" human speech yet), this approach is the best defense for automated attacks in such applications.

Besides automated attacks, security analysis of biometric systems may also consider *human attacks* i.e., the faking of biometric samples for a user u by another user u'. Unlike automated attacks, these attacks (if shown to be feasible) seem harder to defend against (particularly, in remote authentication scenarios) and liveness detection is unlikely to work against them. Some researchers question the feasibility of such attacks based on the position that they require specialized skills [2] and finding skilled people is expensive. Recent work has demonstrated that this position does not hold up for some biometric forms like keystroke dynamics [17] but this work only applies to biometrics for which the notion of a "match" (and particularly, "closeness" of a match between two samples and their temporally-corresponding parts) is visually representable to human attackers. This assumption does not hold for all biometric forms, including voice biometrics, which make limited use of temporal data in creating biometric templates. Furthermore, while [17] studies the question of designing appropriate feedback mechanisms to *train* unskilled users in biometric mimicry, we consider the question of *finding* appropriate mimics in a large universe (e.g., an online crowdsourcing platform) in a manner such that they can succeed with minimal training. We expect that this approach will apply to a broader class of biometric systems and investigate it for voice in the current paper.

[1]Nuance's FreeSpeech system implements this technique: `http://www.nuance.com/landing-pages/products/voicebiometrics/freespeech.asp`

2.2 Speaker Verification Primer

Before we describe relevant literature on security of speaker verification (SV), we provide an overview of SV methods. Broadly, there are two types of speaker verification systems—*text-dependent* [12], which require users' training and test samples to have the same (or similar) text; and *text-independent*, which do not have such a requirement. Both types have multiple real-world deployments, but text-independent systems are gaining popularity because they tend to offer relatively better usability (no human memory requirements) as well as security (greater amenability to liveness detection) trade-offs. At the same time, text-independent techniques are harder to implement and less efficient: unlike their counterpart, they cannot rely on temporal relations between speech frames when modeling speakers and have to work harder to extract features from speech. We focus on text-independent systems in this paper although our method could equally well be applied to text-dependent ones.

Most text-independent SV systems work as follows. To process any input speech, they first create its frequency spectrum (using one of many variants of the Fourier transform) and based on certain properties of the spectrum, extract, what are called, *spectral features* from it. These features are generated by averaging out values across the entire length of the sample i.e. they do not contain temporal data. Spectral features extracted from the training data could either directly be mapped to a biometric template or, what is more common, a *generative model* is learnt over them. Standard machine learning approaches like expectation maximization (EM) are applied to learn such models. The most commonly-used generative models are *Gaussian Mixture Models (GMMs)* which represent speech features in the form of a collection of Gaussian distributions. The process of matching a test speech sample γ to a speaker u involves extracting spectral features from γ and testing the likelihood of these features being generated from the GMM linked with u. Some systems also try to model prosody in speech when representing users but the use of spectral features is more common. We refer the reader to [13] for a good overview of the text-independent SV literature.

In this paper, we focus entirely on one kind of SV scheme—the GMM Universal Background Model (GMM-UBM) scheme [20]—which is the most widely-studied, and possibly, the most widely-deployed, text-independent SV scheme. The key characteristic of this scheme is the use of a "background" model which is meant to model the universe of all human speech and is a GMM, say Λ_B, trained prior to creating speaker models using samples from outside the target set.

The speaker model of a user u, say Λ_u, is then built by "adapting" the background model Λ_B based on features extracted from u's training samples. Matching a sample γ to u involves comparing the likelihood that γ's features were generated from Λ_u and the likelihood that they were generated from Λ_B. A high match score is assigned to γ if the former likelihood is much greater than the latter and the sample is accepted as u's sample if and only if the match score exceeds a pre-set threshold. In UBM-based systems, the better the quality of the background model (more variety in background speech samples), the better is the performance of the system. Besides GMM-UBM, there is a variety of other GMM-based schemes in the speaker recognition literature and some of the more recently-developed ones also provide greater resilience to session variability than GMM-UBM. But these schemes are less standardized (in terms of parameter settings) and stable, well-documented implementations for academic research are not widely available.

In general, there seems to be an upward trend in the adoption and deployment of SV systems worldwide [8], although rigorous data on this is missing. Multiple banks (e.g., bank Leumi in Israel[2]) and telecom operators (e.g., Bell Canada in Canada[3] and Turkcell in Turkey [3]) have already deployed SV systems in their phone-based support services and banks elsewhere in the world are also moving in that direction [25][4]. Conceivably, a good number of these systems are text-independent [3] although accurate penetration statistics are hard to find. In India, we are aware of one company [23] which supplies voice biometric technology for on-site authentication to a large BPO with over 100K customers and has also piloted their technology with multiple financial service providers; one of our future goals is to study usability-security trade-offs in SV systems in collaboration with this company.

2.3 Security of SV Systems

As with other types of biometrics, the literature has largely focused on automated attacks when analyzing speaker verification security. Several papers analyze susceptibility of SV systems against replay and conversion attacks [6, 10, 14] but there is no evidence that these attacks work against the liveness detection measures that have been proposed for voice biometrics. In particular, human mediation and challenge-response seem sufficient to defeat them.

There is prior work on imitation attacks, too, but most of this work is either restricted to studying mimicry of celebrity voices [26] or mimicry performed by professional or semi-professional imitators [1, 15] or else, a combination of the two [10, 11, 26]. The general picture portrayed by these works is that mimicry specialists are good at imitating prosodic elements of speech but tend to perform poorly (false acceptance rates (FARs) of 10% or less) when trying to attack GMM-based SV systems. The work of Lau *et al.* [15] is the only one we are aware of which reports FARs of greater than 30%, but they too seem to consider "amateur imitators" (two in number) with some experience in mimicry[5]. Our work significantly expands the space of amateurs through the use of Web-based crowdsourcing and we incorporate people without any experience in drama or mimicry to play the role of impostors. Prior studies [10, 11, 15, 26] use at most six potential imitators whereas we consider nearly two hundred and carefully narrow down to the most promising candidates from this set. Despite our relatively low-skilled sample space, we are able to find users who can perform successful imitation attacks on SV systems and often with performance better than what has been demonstrated for the case of experienced imitators.

3. THE ATTACK METHOD

Throughout the paper, we assume text-independent SV systems implemented over cellular networks (i.e., we assume all voice communication happens using mobile phones). While this assumption is not necessary for the implementation of our method, it arguably leads to the most convenient implementation of it. Authentication over mobiles forms one of the most compelling application scenarios for speaker verification and many real deployments operate in this scenario.

We now describe our method at an abstract level. Let $\mathcal{A}$ be the SV system being analyzed and let S be the speaker set for which the system is trained. Our attack method involves setting up a telephony server which runs an IVR system for voice data collection. The attack occurs in three steps:

1. *Imitator solicitation:* First, we use a crowdsourcing platform $\mathcal{P}$ to solicit candidate imitators for speakers in S. Workers associated with $\mathcal{P}$ are asked to perform two tasks: (a) submit natural (i.e. unmodified) speech samples to the telephony server and (b) given recorded speech samples of speakers in S, listen to these samples, select some speakers who the worker believes he can feasibly copy and submit "mimicked" speech samples for each selected speaker. We assume an IVR interface which allows workers to listen to their recordings and to re-submit a sample, if the worker perceives a previous recording to be unsuitable. Suitable incentive and disincentive schemes can be used with $\mathcal{P}$ to attract workers to these tasks.

   The mimicry task is meant to identify imitators based on their own judgement of which speakers they are capable of mimicking and their perceived similarity with such speakers. There may be few people who possess the skill to make such judgements accurately but in a large crowd of workers, finding such people is not an impossible outcome. Note that we also collect natural samples per worker, which enables us to match workers to speakers in S based on natural closeness in voice.

[2] `http://www.businesswire.com/news/home/20100415005768/en/Top-3-Israeli-Banks-Roll-Customer-Facing`

[3] IBM's 2012 Case Study titled *World's Largest Voice Authentication Deployment Makes Privacy Protection More Convenient for Bell Customers* discusses this deployment: `http://www-304.ibm.com/partnerworld/gsd/showimage.do?id=24252`

[4] `http://www.biometricupdate.com/201301/mobile-devices-to-drive-bank-adoption-of-voice-biometrics`

[5] The definition of "amateur imitators" is ambiguous in [15]. Based on communication with the authors, it seems that these imitators were less experienced than those used in prior works [10, 11] but it is unclear whether they had prior mimicry experiences or not. FARs from [15] are higher than those from other studies plausibly because the imitators were matched to targets selectively (based on voice similarity) before FAR-computation; however, the study did not use candidate filtering techniques to identify good mimics, the way we do in the current work.

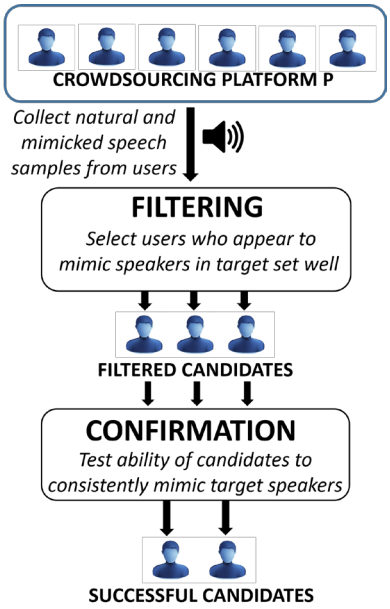


Figure 1: Pictoral depiction of our attack method.

2. *Candidate filtering:* Of all the workers who participate in the above crowdsourcing tasks, we select a few *candidate* imitators based on their performance on these tasks. For each worker w who participates, we determine whether w is a candidate imitator or not using two tests: *(a) do w's mimicked samples for any speaker $u \in S$ successfully authenticate u to $\mathcal{A}$?* and *(b) do w's natural or mimicked speech samples successfully authenticate u' to $\mathcal{A}$ for some user $u' \in S$ (not necessarily a speaker attempted to be mimicked by w)?* A worker is declared a candidate imitator if either of the tests return true for him. If he satisfies the first condition, we refer to him as a *deliberate candidate*; if he satisfies the second one, we call him a *emergent candidate*. Both conditions involve black-box invocation of the test procedure of $\mathcal{A}$. (Since the system is assumed to be text-independent, it is reasonable to test for the second condition using it.) For each condition, different implementations based on different notions of "success" can be used. For example, one implementation of type-2 candidacy testing could be: for any n natural speech samples uttered by w, do at least $n/2$ samples authenticate u' to $\mathcal{A}$ for some $u' \in S$?

   Our objective in including the second test for candidacy is to account for the potential incapability of workers to select good targets and the possibility of a "natural" match between a worker's voice—or mimicked variants of it—and a target's voice. Our technique could be generalized to capture other types of voice variations from each worker (e.g., "fake your voice by raising your pitch") and use the collective information from all variations to decide a worker's ability to mimic users in S, instead of relying only on his mimicry attempts. We restricted ourselves to the above approach for simplicity and even with this approach were able to achieve some success.

3. *Confirmation:* In this step, we try to increase our confidence in candidate imitators being good imitators. For each candidate imitator w identified above and a corresponding matching speaker u, we invite w to perform the following task: listen to the speech samples of u and submit multiple mimicked samples for that speaker. As the worker performs the task, he may also be given instantaneous feedback about his performance in order to help him create future samples better. We evaluate imitators based on their ability to successfully authenticate u to $\mathcal{A}$ in this task multiple times.

In a real implementation, there is also a fourth step in which the adversary selects the top performers in the confirmation step and has them authenticate as their corresponding speakers directly to $\mathcal{A}$. In this paper, we ignore that step since our goal is only to understand attack possibility, not in mounting an attack on a real system.

The assumption about black-box access to the attacked system $\mathcal{A}$ has some advantages. First, it makes the attack simple to implement and powerful from the perspective of proving negative results. (Insecurity against a black-box attacker implies insecurity against arbitrary attackers.) Second, it leads to a generic approach to security analysis; so, for example, the exact same technique can be applied to a different implementation of $\mathcal{A}$ with no change in the individual steps. Finally, it models the real possibility that the adversary may not have enough information about system implementation, and still be interested in breaking it. In practice, there may be limits on the number of black-box calls the adversary can make to the system (which could affect attack efficiency) but it is conceivable that the adversary can "simulate" such black-box access using other means (e.g., by computing matches on an identical copy of the system available as, say, commercial software or by working with a different system but one based on a similar algorithm). Future work is needed to determine how feasible black-box simulation is for real systems.

4. EXPERIMENTAL SETUP

This section presents the experimental setup we used to analyze our attack technique. We used an Asterisk-based IVR server[6] for all our speech data collection from users. Experiments were conducted from a laboratory in Bangalore, India and we chose to use Indian voices for both speakers and impostors in order to ease communication with users.

4.1 Speech Materials Used

While there are many standard speech datasets available for conducting speaker recognition experiments (e.g., the NIST SRE datasets which are updated on a regular basis), there are none with Indian voices that we have found available for free, which is why we decided to create our own dataset[7]. Our target set S consisted of 53 male users employed in a Bangalore-based IT company. Each speaker

[6] Asterisk is an open-source platform for building voice-based applications: `http://www.asterisk.org`

[7] Using standard datasets would have introduced effects of accent mismatch in the mimic selection process which we wished to avoid. Besides, such datasets are available under restricted usage licenses (e.g., use only for evaluating certain new SV techniques), which didn't fit our experiment goals.

provided two training samples (20-30 secs each) and multiple test samples (4-10 sec each) containing a combination of spoken digits and English sentences. Training and test samples were not phonetically matched, although test samples had some repetitions. Across speakers, training (resp. test) samples contained identical text, modulo some minor differences based on speaker identity (e.g., samples contained the name and occupation of the speaker). All speakers provided informed consent for using their speech data for our experiments. Our target set is admittedly small, but this only helps us strengthen our claims regarding the possibility of crowdsourced attacks on SV systems.

Speech was recorded via calls made to our IVR system from one out of two experimental handsets. Speakers spoke in a laboratory environment with limited ambient noise (modulo the sound of air conditioners and PCs). We spent about 5 minutes collecting speech samples per speaker. We focused on male speakers because we expected the task of finding male imitators to be easier than that for female ones (most Indian crowd-workers are male [21]). As our work is a proof of concept for crowdsourced attacks, focussing on males is sufficient to establish the viability of such attacks. Future work is needed to extend our results to female speakers. Our dataset is freely available upon request to the authors.

4.2 SV Settings

For our experiments, we used an implementation of GMM-UBM in the open-source package Alize [5], which is the only open-source package for speaker recognition with an active developer community today. The system was set up to operate on spectral features, as is standard in the GMM-UBM method. Waveforms were sampled at a frequency of 8 KHz and processed in 20ms frames with intervals of 10ms. The feature set consisted of 16th order mel frequency cepstral coefficients (MFCCs) and a log-energy term, augmented with corresponding first order derivatives, to result in a $(16+1)\times 2 = 34$ dimensional feature vector per frame. Standard normalization and energy filtering techniques were deployed to fine-tune the features.

For training the background model, we used a set of 424 speech samples (all male Indian voices) obtained from a set of crowdsourced data collection tasks posted on Amazon's Mechanical Turk (MTurk). These samples were conceivably submitted using a variety of mobile handsets, which would imply a good degree of variance and hence representativeness of the background model. Speaker models for all speakers in S were trained using a maximum a posteriori (MAP) based trainer. As is common in SV implementations, GMMs with 512 distributions were used and for computing test scores, we used average log-likelihood ratios (computed for the "top ten" distributions in the speaker models), refined with a standard normalization technique (T-Norm). We refer to the resulting SV system as $\mathcal{A}$ in what follows.

We evaluated the system's performance using a set of $10s$ test samples, one sample per speaker. (The same samples were used for score normalization across speakers.) Even with relatively short input speech, we recorded a small Equal Error Rate (EER)[8] of 2.31% for our 53-speaker dataset, not a surprising finding given that our data collection took place in a controlled environment. The Detection Error Trade-off (DET) plot for our baseline setup is shown in figure 2. The EER threshold score was determined to be $t_e = 2.04$. In our experiments, we used this threshold to determine success of matching test speech samples to speaker models: when we say that a sample γ *authenticates* u *to* $\mathcal{A}$, we mean that the result of matching γ to the speaker model of u using $\mathcal{A}$ produces a score greater than t_e. We assume that $\mathcal{A}$ also provides an interface to query for the *closest match* to a given test sample γ i.e. the speaker label u for which $\mathcal{A}$'s matching procedure produces the highest matching score between γ and u's speaker model, when compared across models. This interface is required in our attack implementation below. The interface may not exist for a real SV deployment, but it is generally possible for an attacker to simulate it—via standard log-likelihood computations—once he has acquired speech samples from the target speakers. (This is not an onerous task for a determined attacker given that speech is a frequently revealed biometric across users.) Finally, in our descriptions below, when we say that a sample γ *strongly* authenticates u to $\mathcal{A}$, we mean that it authenticates u to $\mathcal{A}$ and u is the closest match to γ in S.

In our settings, we did not make an attempt to optimize the parameters to get the best EER value, and chose standard parameters that are recommended by the literature on Alize. Since our interest is mainly in understanding the *relative* performance of imitators (compared with the baseline), this optimization does not seem necessary.

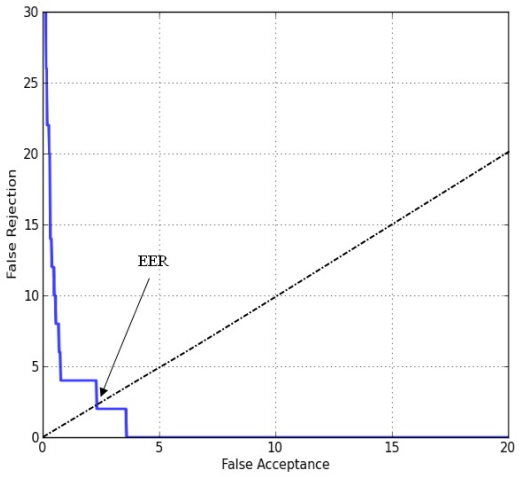


Figure 2: DET plot for our baseline SV system. The EER is 2.31% and the corresponding threshold is 2.04.

4.3 Crowdsourcing Apparatus

We implemented our attack technique using tasks posted on Amazon's MTurk platform. We randomly selected 20 speakers from our dataset and published their test and training samples via a Web interface[9]. Workers were asked to do

[8] Based on the threshold t set for match scores by the system's decision-making procedure, two types of errors can arise: *false rejection rate (FRR)*, the fraction of legitimate test-sample/speaker-model pairs which fail to score greater than t; & *false acceptance rate (FAR)*, the fraction of non-matching test-sample/speaker-model pairs which obtain a score greater than t. The EER is the value of FAR determined for the threshold t_e for which FAR equals FRR.

[9] We chose to publish speech samples of a subset of speakers instead of the entire set in order to ease target selection for our crowd-workers and to ease call management at our end. For studying attack "possibility", this approach is sufficient

the following task: listen to the 20 speakers' samples, select the speakers you want to mimic, practise mimicking them and then call our IVR server from any mobile phone[10]. On each call, the worker was required to provide three speech samples. The first was a natural speech utterance of two sentences which would involve stating their worker ID (a 13-15 character alphanumeric string) and a "task ID" associated with the speaker they are trying to mimic in that call. This was to be followed by two mimicry attempts corresponding to text spoken by the speaker in the published test samples (and lasting 4-10s each in the recordings). Workers could re-record mimicked samples in the same call but only one call per task ID from the same worker was considered for the evaluation. (We still maintained data from the other calls.) We explicitly stated that the task was restricted to male workers in India.

We expected workers to make multiple calls to the system, which would enable us to capture multiple natural speech samples per worker attempting our task. In effect, we recorded, for most workers, at least two mimicked utterances per selected target and two or more natural speech utterances. We emphasized the possibility of a big bonus (5 USD per target) for the best mimicry attempts but provided limited guarantees of payment to bad attempts. We used a subjective, tiered definition for "goodness" of a mimicry attempt. Workers who completed calls as required and who we perceived as making a well-intentioned voice modification (even though unsuccessful) were rewarded with a payout of at least 5 cents for each such call. If the worker made a remarkably good attempt (again, subjectively judged) or if the test scores for at least one of the mimicked sample was at least 50% of the self-scores of the target speaker, the payout ranged between 10 cents to 1 USD, based on our subjective evaluations. The ten best attempts received the bonus in the experiment[11].

Our candidate filtering process was as follows. First, we test whether both of w's mimicked speech samples for a target speaker u authenticate u to $\mathcal{A}$, in which case w is declared a deliberate candidate imitator. The second test, for checking emergent candidacy, involved two sub-tests: we test whether, for some $u' \in S$, either

- a majority of the natural speech samples of w strongly authenticate u' to $\mathcal{A}$; or
- both of w's mimicked speech samples for one target speaker (other than u') strongly authenticate u' to $\mathcal{A}$

If so, we accept him as a candidate. A worker could potentially be both a deliberate and an emergent candidate imitator or satisfy more than one conditions in the emergence test. The motivation for using strong authentication (instead of plain authentication) for the latter test is to increase the likelihood that the observed match between w and a speaker model was not a serendipitous event. All candidate imitators identified in this manner received at least 10 cents each and 5 USD in case they proved to be a deliberate candidate.

and only helps strengthen our results.

[10]While we did not try to rigorously determine the nature of the calling devices, a cursory examination of caller IDs suggests that most callers used mobile phones for the calls.

[11]Some amount of subjectivity in incentivizing workers seems necessary given that mimicry, in general, is judged perceptually and the association between perceptual and quantitative judgements is unclear and opaque to the workers.

Candidate imitators were invited to take part in the confirmation step. In this step, the expectation from the invited workers was to submit more natural speech samples, two per call and at least 60 mimicked samples for the target speaker u that the worker was able to (strongly) authenticate as. Candidates were largely required to utter the same text used in the target test samples but we also collected a few recordings of speech that differed from the test samples in textual content but were similar in the number of constituent syllables. The mimicked samples were collected across multiple recording sessions of at least 20 recordings each. Candidates were promised a bonus of 5 USD per session if they "performed well" in that session, which we defined as authenticating (though not necessarily, strongly so) as u in at least 30% of their attempts. We manually interacted with the imitators during the confirmation task, giving them instantaneous feedback on their performance (over a parallel Google Hangout session) as they submitted fresh recordings to the system and injecting frequent remarks of encouragement. Candidates were instigated to listen to prior "good" recordings of theirs and to try to imitate such recordings, as a strategy to improve scores[12]. In some cases, a target speaker *other than* u emerged as the closest match for the imitator in a majority of the new mimicked recordings; when this happened, we invited the imitator to attempt to mimic the new target afresh. At the end of this interaction, the candidate responded to a questionnaire asking about his background and experience in imitation and on MTurk tasks, in general. Cumulatively, we spent at least 3 hours per candidate during confirmation.

4.4 Mimicry Artists

For the sake of comparison, we also solicited participation in our task from mimicry artists in India. We found contact details of 25 artists (through a human agent who managed their portfolios) and reached out to them by phone. The people we reached were mostly amateurs and enthusiasts of mimicry and their expertise in the art was not well-established. None reported to practise it as their primary occupation although some claimed to have performed imitation acts in competitions, as part of plays and in gatherings.

Our interactions with these artists were similar in nature to those with the MTurk workers except that we interacted over phone more than email, which enabled us to converse with them more openly. The artists used our IVR server to provide sets of natural and mimicked speech samples just like the MTurk workers. We offered incentives similar to those offered to MTurk workers (equivalent of 5 USD bonus for the best attempts) and applied the same filtering techniques. In effect, this part of our experiment was a more targeted form of crowdsourcing aimed at a specific audience which seemed to possess the skills our task demanded.

[12]More elaborate forms of feedback could also be conceived for human attacks on biometric systems. For example, Meng *et al.* [17] provide visual feedback to their imitators of keystroke biometrics on "how close" they are to their target user. Such visual feedback is currently difficult to design for text-independent SV systems because of the lack of temporality in the creation of feature vectors in these systems. (The feedback giver cannot easily depict in a picture which part of the speech is being mimicked well and which part is not.) As such, we restricted ourselves to giving overall score-based feedback to our participants and included oral forms of encouragement along the way.

5. RESULTS

Over the 8-week period that our experiment ran in, we received a total of 733 calls to our IVR server, which included calls from both MTurk workers as well as some of the artists. Out of these, about a hundred calls generated audio files which encountered errors against Alize (either due to IVR bugs or because of missing voice data) and some others had issues of missing information (e.g., missing or incoherent MTurk ID) or were from a female caller. Discarding all such cases, we were left with 493 calls from 180 unique callers—176 MTurk workers and 4 artists. In our analysis, we used data only from these calls.

Persuading the artists to sign up for our task proved more difficult than we expected. There were multiple reasons for this difficulty. Some artists indicated that they had stopped practising the art. Others were not excited about imitating non-celebrity voices. A few felt that our incentives were not sufficient (we adjusted our offering in such cases, though this did not affect the eventual callers). Finally, many indicated an intent to call but never did, conceivably for a lack of real interest or distrust. While this was disappointing in a way, a useful side-effect was that the few artists who did participate were highly motivated to perform our task, as was exhibited in their repeated calls (more than 175 recordings in the case of one artist) during the experiment.

5.1 Candidate filtering outcomes

Among the 176 valid MTurk callers, 39 were determined as candidate imitators for our system—2 deliberate candidates and 38 emergent candidates, with one overlapping the two criteria. When probed further, we learnt that one of the two deliberate mimics had used a record-and-replay technique to impersonate as his target to the system instead of self-imitating the voice[13]. The other performed better during the confirmation phase, as discussed below. Effectively, only one out of 176 MTurk workers emerged as a true deliberate candidate in our experiment.

While the finding of a single deliberate candidate may seem like a dismal outcome, it is remarkable in the light of the fact that workers selected their targets from a sample of size 20 and did not have any visibility into the workings of the SV system being tested. As a comparison, none of the artists qualified as deliberate candidates, an indication that experience in mimicry may not be a criterion for imitation attacks against speaker verification systems. We gave the artists the additional capability of accessing recordings of all speakers in S and mimicking as many as they desired; still, deliberate candidacy proved difficult for them.

Even the finding of emergent candidates is interesting. Given that we had only 53 speakers in our dataset, we find it surprising that over a fifth of the 176 workers being evaluated could match *some* speaker in this set so as to be able to strongly authenticate as that user multiple times. Not all of these workers matched to unique users: the 38 candidates were mapped to 21 target speakers, with one target speaker emerging as the closest match for *six* candidates. Our confirmation task tested the resilience of some of these matches by collecting more samples from the workers.

The artists did slightly better on emergent candidacy, 2 out of 4 (50%) of them satisfying the condition. We enroled a third artist for the confirmation task even though he did not strictly qualify as a candidate. This artist was the most enthusiastic participant amongst all mimics: he attempted mimicry on more than 20 targets, was ostensibly making significant modifications to his voice in his mimicry attempts and even though he failed the emergent criteria on all targets in the first attempt, he returned to make further attempts wherein he was successful in meeting it for one speaker.

Interestingly, most of our emergent candidates were declared emergent not because of a close match between their *natural* voice and a speaker in our dataset, but because of closeness between their *faked* (mimicked) voice and a speaker they didn't intend to mimic. Out of the 38 MTurk emergent candidates, this was true for 28 of them, which included one worker who satisfied both criteria. The same was true for all the three artists who were emergent candidates. This suggests that when evaluating the ability of a user as an imitator for speakers in a system, simply matching his natural voice to the existing speaker models, as done in past research [15], is not sufficient; requiring him to vary his voice may give better hints on who his closest matches could be. The success of some of our emergent candidates, as discussed below, in continuing to impersonate their targets further supports this hypothesis.

Overall, our key learnings from this part of the experiment were: (a) most MTurk users do not have the ability to *self-identify* which speakers from a given dataset they can mimic to an SV system, but people experienced in mimicry do not seem to possess that ability either (within the scope of "ordinary" non-celebrity voices); and (b) even though such an ability may be scarce, several users (workers as well as artists) may be able to create voice modifications which bring them unexpectedly close to such speakers.

5.2 Confirmation outcomes

Out of the 38 MTurk workers we invited to participate in our confirmation task, 13 (i.e. 34%) responded with an affirmative response. We sent multiple follow-ups to the non-respondents but this number did not change. Even amongst the respondents, 4 out of the 13 responded only after multiple invites. (See the appendix for the email template we used.) While there could have been several reasons for this poor response rate, we believe that the peculiar nature of our tasks (expectation of mimicking others' voices, doing it over phone and doing it multiple times) influenced participant behavior and likely raised a sense of suspicion or distrust amongst the workers who did not respond. Some of the respondents even expressed concern in their initial email responses, one going as far as saying:

> *I got a little concern[ed] about my privacy when going through [your email]. Can I know [what] you need the recordings for?*

Nevertheless, the thirteen workers we recruited provided us with sufficient data to demonstrate the possibility of our attack technique being successful and we did not attempt further solicitation of workers, nor did we try too hard to allay potential feelings of distrust. In real attacks, it is unlikely that the attacker would use a platform like MTurk, relying instead on more targeted platforms (with better support for subversive activities) to conduct the attack.

The artists were more responsive than the workers (all

[13] Only 2 workers tried this technique to fool our system, out of which 1 passed our filtering criteria, a plausible indication that malicious intent is scarce amongst MTurk users.

three candidates completed the confirmation task) plausibly because our engagement with them was less anonymous and more conversational in nature, which may have increased trust in the activity. Overall, 16 candidate mimics attempted our confirmation task—13 workers, 3 artists—which is more than the number of mimics used in any prior research on imitation attacks [11].

Measure	MTurk	Artists
#(participants solicited)	> 200	25
#(participants who made valid calls)	176	4
#(candidate imitators filtered)	38	3
#(candidates who replied to emails)	13	3
#(candidates who were successful)	6	3

Table 1: Summary of our filtering and confirmation outcomes. We don't have data on the exact number of MTurk workers solicited but based on the calls received, it is clear that there were more than 200 of them. Through our iterative filtering process, six of these were confirmed to be successful imitators at the end. Of the 25 solicited artists, 3 were confirmed as being successful.

In post-hoc interviews, none of the MTurk workers reported to have had any training in mimicry or drama in the past although four of them claimed to have practised casual mimicry in the company of friends and family. The artists were more experienced, but not significantly so, with one artist reporting not to have done stage performances ever and another reporting to have had extensive voice-over experience but none so in celebrity mimicry. Geographically, these individuals were dispersed across India with exactly half of them from the South and the remaining from northern India. Their ages ranged from 20 to 63 (median age of 26) and their personal incomes varied from 33 USD to 1100 USD per month (median income of 42 USD per month)[14].

Our confirmation procedure ran for a cumulative period of six weeks and we collected a total of 1060 speech samples from our candidate imitators during this period. Each candidate mimic called from a mobile phone but the phone model differed across candidates. We did not attempt to control for recording environment except for a general advice to call from a quiet room. In our analysis of the confirmation data below, we use the actual scores of candidate imitators against speaker models used by our SV system $\mathcal{A}$ and not just the binary outcome of its matching procedure. This is done purely for the sake of analysis and does not affect attack implementation; a real-world adversary may not have access to scores from the SV system if it is assumed to be black-box accessible only.

5.2.1 EER-based Evaluation

Overall, the performance of our candidate imitators declined in the confirmation stage but a majority of them continued to authenticate their associated targets to the system across sessions. For each candidate w, we computed his individual false acceptance rate (FAR)—the number of w's speech samples that could authenticate his target speaker to the system (at the EER threshold) divided by the total number of speech samples evaluated for him. The FARs for the nine leading candidates were observed to be consistently over 20% across sessions. For the remaining seven, we observed FARs of less than 20% in the first two sessions and for the most part, we did not engage them beyond the second session. Our analysis below focuses on the 9 leading candidates. We refer to the workers amongst these as $\mathsf{w}_1, \ldots, \mathsf{w}_6$ and the artists as $\mathsf{a}_1, \ldots, \mathsf{a}_3$.

Each of these candidates participated in 3 to 5 well-separated recording sessions (inter-session separation of at least a day) of at least 20 recordings each. Each participated in at least two contiguous sessions with individual FARs exceeding 0.3 and we continued recording until this was accomplished by each of them (going beyond the third session where required). The mean FARs for the candidates in their *last 3 sessions* are depicted in Table 2 (in the column labeled FAR). The total number of confirmation sessions conducted per candidate is denoted n. Out of the 9 candidates, only one, namely w_1, is a deliberate candidate and of the remaining, only two (marked with a superscript nat) were identified based on the closeness of their *natural* voice to that of the target. The remaining emerged candidates due to an observed closeness between their faked (mimicked) voice and their target's voice.

Source	Label	Type	n	FAR	modifier?
MTurk	w_1	deliberate	5	0.424	N
	w_2	emergent	5	0.683	Y^+
	w_3	emergent	4	0.417	Y^+
	w_4	emergent	3	0.417	N^+
	w_5	$\text{emergent}^{\mathsf{nat}}$	3	0.367	N
	w_6	$\text{emergent}^{\mathsf{nat}}$	3	0.333	N
Artists	a_1	emergent	3	0.567	Y^+
	a_2	emergent	3	0.383	Y^+
	a_3	emergent	3	0.533	Y^+

Table 2: Overall False Acceptance Rates of the nine leading candidate imitators.

All of these nine candidates were able to authenticate as their target speakers in at least 33% of their (last 60) recordings. This is significantly greater than the 2.3% FAR rate that the system was initially calibrated for and it suffices to launch online attacks on a real system. Averaged across all candidate imitators and all speech data used in the table, we computed an FAR of 45.8% and for the six MTurk workers, this figure stood at 44%. The finding is made all the more significant by the fact that we did not control for environmental and channel effects in the voice recordings; the imitators' handsets and speaking environment could have been very different from that of the speakers, which could have made it hard for them to match the speaker voices.

The target speakers associated with these candidates are not all unique: w_4 and a_3 share the same target and so do w_6 and a_1. The target for w_1 is shared by 5 other candidate imitators, although only w_1 reached the confirmation stage of our experiment (the rest did not respond to our invitations).

The last column in the table (labeled modifier?) depicts whether a majority of a imitator's mimicked speech samples were seen to match the target speaker's voice more closely than his natural voice—a likely indication that the imitator was making a significant effort to modify his voice to match the target. This metric was computed by first creating speaker models for all the nine candidates in the system and then invoking the matching procedure on the imitator

[14] We use a USD to INR conversion rate of 1:60 for these computations.

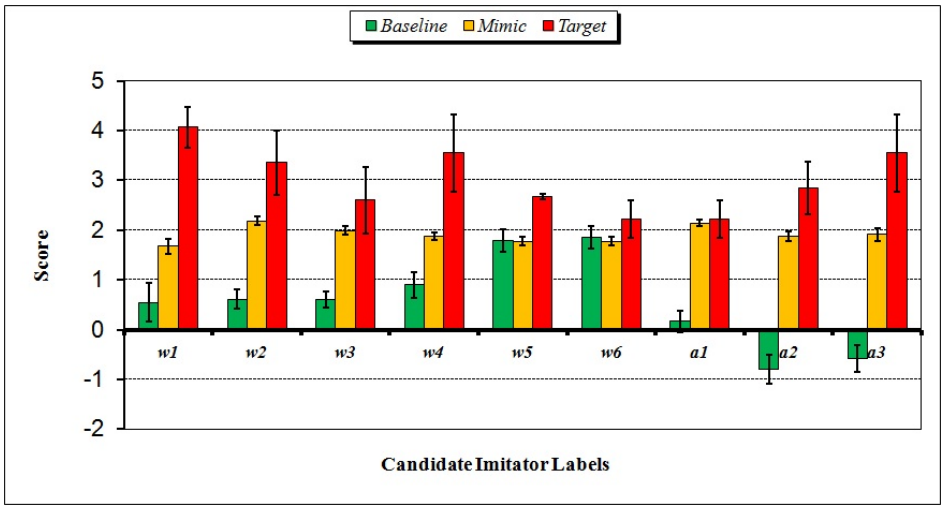


Figure 3: Comparison of candidate imitator scores with target self-scores: *Baseline* refers to the score of an imitator's natural speech samples against the target speaker model and *Mimic* is the score of his mimicked speech samples. *Target* is the target speaker's self-score. Error bars depict standard error of mean.

speech samples for both the target speaker model and the imitator speaker model. Two of the MTurk candidates and all of the artists were found to be modifiers by this definition. A superscript of + indicates candidates who reported to have had mimicry experience in the past, which was true for workers w_2, w_3, w_4 and all the artists. (This coincides almost perfectly with our modifier set.) Notice that the modifiers are also better performers on average: the mean FAR for the modifiers is 0.52, whereas that for the rest is 0.39. While mimicry skill or experience does not seem necessary for launching successful imitation attacks on SV systems (as demonstrated by the performance of non-modifiers in our set), it does seem to aid attack success.

In terms of inter-session variability in mimicry performance, our data does not reveal a consistent trend across imitators—some (in particular, w_1, w_2 and a_3) improved with time while others performed non-monotonically although per-session FARs remained consistently above 20%. Intra-session trends are also not monotonic and in particular, success likelihood did not necessarily increase across attempts within a session. These findings could be explained by two characteristics of our experiment: first, effects of learning could have been negated by factors like boredom and fatigue within sessions; and second, as we find below, match scores for most candidates are generally in the neighborhood of the EER threshold and could thus be more sensitive to inter- and intra-session changes than the target voices. Future work is needed to address these issues and in particular, to devise feedback mechanisms that have sustained effects on mimic performance. As indicated earlier, this is challenging for text-independent voice biometric systems because of the lack of temporality in feature creation in such systems.

Only one imitator (w_4) was asked to change his target speaker in the process of confirmation (because of greater closeness observed with that target during initial authentication attempts); the rest mimicked their initially-assigned targets. Two of the artists (namely, a_1 and a_2) attempted mimicking multiple speakers based on target suggestions provided by us but their FARs were smaller (less than 10%) for all such targets, an indication that *consistently* and *successfully* mimicking *multiple* targets is a difficult undertaking, even for people experienced in mimicry.

The gap between the nine successful candidates and the remaining seven who took part in the confirmation stage was striking. The latter provided a total of 279 speech samples across 1-2 sessions each but achieved an average FAR of only 4.9% for their respective target speakers. Individual FARs for these seven candidates ranged between 0 and 0.125 with a median of 0.044. These findings highlight the importance of including a confirmation step in the attack as a way to weed out serendipitous matches during candidate filtering.

5.2.2 Comparison with self-scores

Although EER-based analysis gives us some indication of candidate performance, by itself it does not present a complete picture of attacker capability: even by crossing the EER threshold repeatedly, an imitator could be far from the expected target score, something that an SV system could be programmed to detect. As such, it is also useful to compare the scores of these imitators against the *self-scores* of the target speakers i.e., the scores computed on their test samples against their speaker models.

For the comparison, we used test samples we initially collected from each speaker in S, half of which had deliberate background noise (sounds from a busy Indian street) included in the recording. We did not attempt to exaggerate the noise addition and even with its presence, all nine targets' scores were above the EER threshold, on average. Incorporating noisy test samples in the analysis reduces the challenge for the adversary, but also models a more realistic scenario: test samples for honest users are unlikely to be all clean in reality, but an attacker can control his test environmental conditions (e.g., avoid calling from the street). We compared three statistics: candidates' natural speech sample scores against target speaker model (*baseline*), candidates' mimicked speech sample scores against target speaker model (*mimic*) and the targets' self-scores (*target*). Results are shown in figure 3. We use unpaired, 2-tailed t-tests for

measuring significance in our statistical analyses below, with a threshold p value of 0.01.[15]

It is clear that most candidates are modifying their voice in order to mimic their target speaker, although the artists seem to be making more significant modifications. The results for artists a_2 and a_3 are particularly striking—even with natural voices whose match against the target have opposite polarity as the target's self-match, these artists were able to obtain scores that are in the proximity of the target self-scores. The MTurk workers, being relatively less skilled, are making limited modifications and seem to be relying on their inherent closeness with the target's voice. (This is consistent with prior literature on the ability of skilled imitators vs. ordinary humans in being able to modify their voices both in terms of spectral features and prosody [26].) In particular, for workers $\mathsf{w}_1, \mathsf{w}_5, \mathsf{w}_6$, the candidates claimed to be least experienced in mimicry, the difference between the natural speech scores and mimicked speech scores is statistically insignificant.

Second, even though most candidates are making noticeable jumps in moving from the baseline to the mimic conditions, their mean mimic scores never exceed their targets' self-scores, although the difference is statistically insignificant for 6 out of our 9 leading candidates, namely, workers $\mathsf{w}_2, \mathsf{w}_3, \mathsf{w}_6$ and all the artists.[16] Artist a_1, in particular, exceeds his target's mean self-score in more than 40% of his attempts. The performance of artists observed in our experiments surpasses that of mimicry specialists used in prior works like [10, 11, 15, 26], which did not deploy candidate filtering techniques like ours in selecting their mimics (instead relying purely on perceptual mimicry ability)[17]. Finally, we note that even though the workers' overall performance is poorer than that of the artists in our experiment, they compare favorably with that of the latter both in terms of EER-based measures (table 2) and in terms of the means of the raw scores (figure 3); workers $\mathsf{w}_2, \mathsf{w}_3$ and w_4, in fact, surpass a_2 on both these measures and also surpass artists used in prior works [10, 11, 26].

Overall, we learnt three key things from this part of our analysis: (a) most candidate imitators identified by our filtering techniques exhibit good mimicking capability in confirmation tests although the artists are more consistent than the MTurk workers; (b) in terms of absolute scores (EER-based evaluation), these candidates present a potent threat to the system but when viewed relative to target self-matches, they seem less competent; and (c) it is possible that *some* imitators picked from MTurk (like $\mathsf{w}_2, \mathsf{w}_3, \mathsf{w}_4$) can surpass more experienced ones (like a_2) as SV-impostors in absolute measures (which is a new finding relative to the literature) but their *overall* ability to imitate others seems less than that of the latter (which is consistent with the literature [26]).

[15] Although the *baseline* and *mimic* distributions are not strictly independent (same speaker for both), it is conceivable that the speaker applies independent techniques in generating them, which justifies the use of the unpaired test.

[16] In parallel experiments, we have also tested the performance of different forms of record-and-replay attacks on the same system and found them to be able to match, and often exceed, target self-scores for a large number of users.

[17] Most prior work except [11] does not use self-score comparisons, the way we do, which makes comparing against such works difficult. However, even based on EER-based evaluations, our results seem stronger.

6. DISCUSSION AND CONCLUSION

Prior work has contended that human attacks on biometric systems are possible only by people with skill or expertise to imitate others [2, 26]. Our work shows that this is not necessarily the case and it is the first that does so for voice biometrics. Even with a relatively small target set of fifty-three speakers, we were able to find ordinary, untrained people on an online crowdsourcing platform with the capability to impersonate (in absolute EER-based measures) some of these speakers to a well-studied SV scheme. We believe that the geographic and cultural spread of MTurk was critical in enabling us to reach this result: hand-picking candidate imitators from our vicinity may not have been as successful, at least in discovering a deliberate candidate like w_1 as we did in our experiment. Furthermore, the strategy of matching imitators to the right targets based on the formers' faked voices rather than their natural ones seems to have helped—most of the emergent candidates in our experiment were matched to their targets via this approach.

Even for people with experience (artists or professionals), our work provides an improvement over what has been shown in the literature till now. Our artist imitators were recruited after deliberately contacting 25 individuals over phone and carefully mapping them to "close" targets; in the process, we may have ended up selecting some of the more intrinsically-motivated individuals than others did in their mimicry experiments [10, 11, 26]. Our mimicry artists not only exhibited good FARs with respect to their target speakers, but they did excellently in terms of matching target self-scores as well, which is better than what prior work reports (e.g., the professional imitator used in [11] managed an FAR of only about 10% and did not match any of his target's self-scores).

The observation that MTurk workers could not surpass their targets' mean self-scores on average (even after diluting the latter with ambient noise) is our main negative finding from the perspective of attackers. A potential defense against MTurk-based imitators could thus simply be to require every user's test sample to match his or her prior test sample scores in expectation (or do this for the more vulnerable users identified from the analysis). However, not all systems may be in a position to implement this defense for their users, especially in order to be able to handle unexpected session variabilities. To the extent that there remain SV systems with EER-based decision procedures in deployment, the threat from crowdsourcing-based imitation attacks to these systems will also remain.

The overall performance of MTurk workers was worse than of the artists in our experiments and the fraction of successful workers was also considerably smaller (only 6 out of the 176 valid callers i.e. about 3%, succeeded in the EER-based evaluation). But these findings should be viewed in the light of the fact that MTurk workers are easier to find and cheaper to recruit than typical artists and professional imitators. The lower success rate of the workers could potentially be compensated for by expanding the sample space of crowd workers (e.g., assuming a rate of 3% for finding successful mimics, adjust the sample space size based on the number of successful mimics required[18]). Adjusting the

[18] We caution that the sample space and the target set used in our experiments are small; more experiments are needed to confirm the rate of finding successful mimics on MTurk.

sample space for professional mimics is harder because they are difficult to find in the first place.

The fact that six of our workers' performance came so close to that of the artists, and exceeded the latter in a few cases, encourages the continued usage of online crowdsourcing for large-scale impostor search. Future work is needed to understand how best to set incentives for online crowd workers so as to attract more of them to complete such tasks without compromising on work quality. Future work is also needed to understand how varying authentication criteria (like requiring phonetically-richer test samples from users) or the biometric modeling process can affect the performance of candidate imitators discovered by our method. We believe that increasing the phonetic complexity of either the test samples or training samples (or both) is likely to increase resistance to mimicry attacks, but note that this also affects usability for honest users. Achieving imitation-resistance while maintaining system usability is where the challenge lies.

6.1 Implications for Real Systems

Although we have demonstrated the possibility of crowdsourcing-based attacks on SV schemes, the *feasibility* of these attacks and the scale at which they can be mounted on real systems is still unresolved. Will the most capable imitators discovered using the method be able to successfully impersonate their target as they converse with a real system that implements liveness detection? Is it likely that "many" imitators can be found to do this? And, most importantly, are there crowdsourcing platforms where it is possible to find sufficiently many people with the motivation to help break a real system? Recent work [24] shows that for some types of malicious objectives like online vandalism and fake account creations, systematic use of crowdsourcing platforms has already evolved but to the best of our knowledge, such practice is not yet prevalent for attacking biometric (or other forms of) authentication yet. The attack analysis presented in this paper suggests that when this practice shapes up, the resulting attacks are also likely to be quite successful.

While it is important to extend our study to understand attack feasibility, we believe that the most immediate implication of our work to real systems is that it gives system developers a new tool to *analyze* the security of their systems with. Using our method, they can simulate imitation attacks on their systems more easily and, arguably, more cheaply than they could by hiring mimicry artists, which we experienced to be an excruciating and slow process in our study. It also helps them get a better perspective on which speakers in their dataset are more vulnerable to imitation attacks (the so-called "lambs" in the system [9]) than they would by trying such attacks with a handful of professional mimics or considering within-dataset impostors only. For example, in our own attack implementation we observed one target speaker emerge as the closest match for six different candidate imitators during the candidate filtering process (Sec. 5.1). Even though we could confirm this closeness with only one candidate (our deliberate candidate, w_1), it is plausible that more of the others would also have proven as consistent imitators with respect to that speaker, had we recruited them for confirmation. Such vulnerability assessment of individual speakers is impossible if one restricts the analysis to speakers within the target speaker-set; in our case, this "lamb" speaker was found *not* to be a consistent closest, or even second-closest, match for any of the speakers in S. Assessing speakers in this manner could inform the customization of system parameters for preventing imitation attacks on individuals in the dataset.

Other biometric forms could potentially also benefit from crowdsourcing the search for impostors, if not now, then at least in the near future. As more people in the developing world go online and join the crowdsourcing workforce, and as sensing devices like fingerprint scanners and cameras become more ubiquitous, new possibilities for large-scale, cheap and efficient biometric data collection will open up. Such data collection and subsequent analysis can lead to new insights on system vulnerabilities as we discovered in the case of voice in our experiment. Of course, the question of feasibly translating such data collection into real attacks will still remain, but independent of it, existing systems can benefit from the search for impostors in crowdsourced data and prepare better for attacks that may occur in the future.

7. REFERENCES

[1] G. Ashour and I. Gath. Characterization of speech during imitation. In *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, pages 1187–1190, 1999.

[2] L. Ballard, F. Monrose, and D. Lopresti. Biometric authentication revisited: Understanding the impact of wolves in sheep's clothing. In *Proc. of Usenix Security*, pages 29–41, 2006.

[3] BBC. Hello, is that really you? `http://www.bbc.com/news/business-24898367`, 2013.

[4] BiometricUpdate.Com. Coursera looks to verify online student identity with photo, keystroke dynamics. `http://www.biometricupdate.com/201301/coursera-looks-to-verify-online-student-identity-with-photo-keystroke-dynamics`, 2013.

[5] J. Bonastre, F. Wils, and S. Meignier. Alize: A Free Toolkit for Speaker Recognition. In *Proc. of ICASSP*, 2005.

[6] J.-F. Bonastre, D. Matrouf, and C. Fredouille. Artificial impostor voice transformation effects on false acceptance rates. In *Proc. of Interspeech*, pages 2053–2056, 2007.

[7] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp. Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading. *Transaction on Image Processing*, 15(10), 2006.

[8] CIO Journal. Banks Eye Voice Biometrics to Verify Customers. `http://blogs.wsj.com/cio/2013/05/09/banks-eye-voice-biometrics-to-verify-customers/`, 2013.

[9] G. R. Doddington, W. Liggett, A. F. Martin, M. Przybocki, and D. A. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, 1998.

[10] M. Farrus, M. Wagner, D. Erro, and J. Hernando. Automatic speaker recognition as a measurement of voice imitation and conversion. *The International Journal of Speech, Language and the Law*, 17(1):119.

[11] R. G. Hautamaki, T. Kinnunen, V. Hautamaki,

T. Leino, and A.-M. Laukkanen. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Proc. of Interspeech*, 2013.

[12] M. Hebert. Text-dependent speaker recognition. In *Springer handbook of speech processing (Heidelberg, 2008)*, pages 743–762. Springer Verlag, 2008.

[13] T. Kinnunen and H. Li. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. *Speech Communication*, 52(1):52.

[14] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li. Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In *Proc. of ICASSP*, 2012.

[15] Y. W. Lau, M. Wagner, and D. Tran. Vulnerability of Speaker Verification to Voice Mimicking. In *Proc. Int. Symp on Intelligent Multimedia, Video and Speech Processing (ISIMP '04)*, pages 145–148, 2004.

[16] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. Impact of Artificial Gummy Fingers on Fingerprint Systems. In *Proc. of SPIE, Optical Security and Counterfeit Deterrence Techniques IV*, volume 4677, pages 275–289, 2002.

[17] T. C. Meng, P. Gupta, and D. Gao. I can be You: Questioning the use of Keystroke Dynamics as Biometrics. In *Proc. of NDSS*, 2013.

[18] Nuance Communications. Private communication, 2013.

[19] Reuters. German group claims to have hacked Apple iPhone fingerprint scanners. `http://www.reuters.com/article/2013/09/23/us-iphone-hackers-idUSBRE98M01X20130923`, 2013.

[20] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, Jan. 2000.

[21] J. Ross, I. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the Crowdworkers? Worker Demographics in Amazon Mechanical Turk. In *CHI Extended Abstracts 2010*, pages 2863–2872, 2010.

[22] B. Toth. Biometric Liveness Detection. *Information Security Bulletin*, 10, 2005.

[23] Uniphore. `http://uniphore.com/`, 2008.

[24] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and Turf: Crowdturfing for Fun and Profit. In *Proc. of WWW*, 2012.

[25] ZDNet. Indian banks explore voice biometrics for added security. `http://www.zdnet.com/in/india-banks-explore-voice-biometrics-for-added-security-7000018883/`, 2013.

[26] E. Zetterholm. Detection of speaker characteristics using voice imitation. In *Speaker Classification II*, LNAI 4441, pages 192–205, 2007.

APPENDIX

A. INVITATION EMAIL SENT TO MTURK WORKERS

Given below is the email template we used to invite MTurk workers identified as candidate imitators to participate in the confirmation task:

Dear MTurk worker ——,

Based on your performance on our HIT, you have been selected to do a bonus task for us in which the minimum pay is 5 USD. In this bonus task you will be required to make at least 20 voice recordings on our system.

Kindly email —— if you are interested in doing this bonus task. Please specify your name and MTurk ID in the email. Based on your response, we will send you more details about the bonus task.

Looking forward to hearing from you!

Understanding and Specifying Social Access Control Lists

Mainack Mondal
MPI-SWS
mainack@mpi-sws.org

Yabing Liu
Northeastern University
ybliu@ccs.neu.edu

Bimal Viswanath
MPI-SWS
bviswana@mpi-sws.org

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

ABSTRACT

Online social network (OSN) users upload millions of pieces of content to share with others every day. While a significant portion of this content is benign (and is typically shared with all friends or all OSN users), there are certain pieces of content that are highly privacy sensitive. Sharing such sensitive content raises significant privacy concerns for users, and it becomes important for the user to protect this content from being exposed to the wrong audience. Today, most OSN services provide fine-grained mechanisms for specifying social access control lists (social ACLs, or SACLs), allowing users to restrict their sensitive content to a select subset of their friends. However, it remains unclear how these SACL mechanisms are used today. To design better privacy management tools for users, we need to first understand the usage and complexity of SACLs specified by users.

In this paper, we present the first large-scale study of fine-grained privacy preferences of over 1,000 users on Facebook, providing us with the first ground-truth information on how users specify SACLs on a social networking service. Overall, we find that a surprisingly large fraction (17.6%) of content is shared with SACLs. However, we also find that the SACL membership shows little correlation with either profile information or social network links; as a result, it is difficult to predict the subset of a user's friends likely to appear in a SACL. On the flip side, we find that SACLs are often reused, suggesting that simply making recent SACLs available to users is likely to significantly reduce the burden of privacy management on users.

1. INTRODUCTION

Online social networks (OSNs) are now a popular way for individuals to connect, communicate, and share content; many of them now serve as the de-facto Internet portal for millions of users. On these sites, users are encouraged to establish friendships and upload content, providing an incentive for users to return. As a result, social network users today have hundreds of friends and many thousands of pieces of content. These same users are also expected to *manage* their privacy—i.e., select the appropriate privacy setting for each piece of content—a task that is both time-consuming and complex [34].

While much OSN content is shared with default settings (e.g., visible to all of a user's friends), certain sensitive content is often shared with subsets of friends. For example, on Facebook, users may explicitly enumerate friends to allow or deny the ability to view a photo, or create friendlists for the same purpose. We refer to these resulting sets of users who are able to access content as *social access control lists* (social ACLs, or SACLs); by definition, a SACL is a proper subset of a user's friends who are selected by the user to access a piece of content. Due to the privacy-sensitive nature of the content the SACLs protect, one of the hardest parts of using today's OSN privacy management tools is defining appropriate SACLs for different pieces of content.

Many prior studies have examined the privacy concerns that arise when users share content on Facebook, such as the problem of "over-sharing" content with default settings that make the content visible to everyone in the network [34]. As a solution, researchers have proposed grouping friends into subgroups based on their relationship type (e.g., high school friends, work colleagues, family) or community structure in the one-hop network of the user, and sharing content with specific subgroups [16]. However, most of these approaches rely on small scale user studies where they conduct a survey to understand the privacy preferences of users to evaluate their technique. None of these approaches have been evaluated on how well they capture real privacy preferences specified using SACLs. Given that content shared with SACLs is likely to be the most privacy sensitive (and therefore, likely the most important), having an understanding of the SACLs in-use is crucial to designing improved privacy mechanisms for OSN users.

In this paper, we make three contributions: *First*, we conduct the first large-scale measurement study of use of SACLs in OSNs. Using a popular Facebook application installed by over 1,000 users, we collect a total of 7,602 unique SACLs specified by users.[1] We find that over 67% of users are sharing at least some of their uploaded content using SACLs, and that 17.6% of all content is shared with a SACL; these observations underscore the important and unstudied role that SACLs play in users' privacy management.

Second, we focus on understanding the membership of SACLs (i.e., how are the friends who are allowed to view

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

[1]Our study was conducted under Northeastern University Institutional Review Board protocol #14-01-09.

a piece of content similar to each other, but different from other friends?). Examining the in-use SACLs that we collected, we find that for less than 10% of SACLs all the members of the SACL share a common profile attribute. Moreover, we find that only 20% of SACLs show strong community structure in the links between their members. Taken together, these results suggest that SACLs are likely to be difficult to detect automatically. This result is surprising given the existing work on automatically grouping friends based on network structure or attributes for better privacy management [16, 39]; We suspect that this difference occurs because these prior studies did not evaluate their techniques against ground-truth data about fine-grained content sharing in OSNs.

Third, we explore the difficulty faced by users in specifying SACLs today. Overall, we find that the complexity of SACLs (as defined by the number of terms[2] a user must select when creating a SACL) is quite high for a non negligible fraction of our users: over 18% of users specify more than 5 terms per SACL on average. We observe that there is significant room for improvement in reducing the burden of specifying SACLs, and we find that simply allowing users to re-use previously used SACLs reduces much of the user overhead: for the vast majority ($> 80\%$) of users, 90% of their content is shared with fewer than 5 unique SACLs.

The remainder of the paper is organized as follows. Section 2 presents background and related work on SACLs and OSN privacy. Section 3 describes how we obtained our SACL data set, and Section 4 provides some high-level statistics on SACLs. Section 5 explores the relationship between SACL members, while Section 6 investigates the user overhead in specifying SACLs today. Finally, Section 7 concludes.

2. BACKGROUND AND RELATED WORK

In this section, we first provide some background on how social networking sites have evolved in helping users to manage their data privacy today. Our focus is on the fine-grained privacy management tools that enable the sharing of privacy sensitive content on OSNs.

In this paper, we focus on the largest social networking site as of March 2014 — Facebook. Up until 2005, Facebook split users on the site into different regional networks (based on geography, workplace or educational institution). By default, each user would share all of her content with everyone in the regional network and the service lacked any concrete privacy controls for sensitive data. By 2009, Facebook had 300 million [15] users and some regional networks grew too large (e.g., in India and China) to be used for privacy settings. There were widespread demands for better privacy management mechanisms for users [7], and by the end of 2009, Facebook rolled out more fine-grained privacy controls.

2.1 Mechanisms for privacy management

In December 2009, Facebook made an important change which allowed users to set access control policies for content they publish on a per-post basis [23]. For example, a user can share a particular photo with only family members and close friends. This change allowed users to customize their privacy settings on a per-content basis, instead of simply adopting the default privacy setting offered by Facebook, which allows access to "everyone" (all users on Facebook) [19].

Facebook introduced an additional mechanism called *friend lists* [26] to complement their existing fine-grained privacy controls. Users can create friend lists and add a subset of their friends to each of these lists. For example, a user can create a list called "co-workers" and manually add all of her friends who are co-workers into that list. This allows the user to group her friends into different lists that might be meaningful to her in terms of sharing content. Now, instead of handpicking individual friends for specifying a privacy setting for each content, users can use their pre-created friend lists for specifying privacy settings (e.g., share this photo with "soccer buddies" list). Friend lists are private to the user who creates them.

By October 2010, Facebook observed that only a small percentage (5%) of Facebook users had ever created friend lists [27]. This could be due to the manual effort required of the users to create and maintain friend lists. To help users further, Facebook started automatically creating friend lists for the user and populated the lists with a specific subset of the user's friends [14]; these lists are called *smart lists*. This automation is done by leveraging the profile attributes of the user and the user's friends, e.g., employer, location, family and education information provided by users. An example would be a list called "Family" that automatically groups all the friends of the user who have marked the user as a family member. In addition, Facebook also creates two empty smart lists for the user, "Close Friends" and "Acquaintances". However, instead of auto-populating these two lists, Facebook only shows friend recommendations to the users based on the interaction between the user and her friends. In this paper, we will refer to all of these Facebook-created smart lists as *Facebook lists*. Moreover, when using the term *lists* we are referring to both the user-created friend lists as well as Facebook lists.

So far, we observed that there are different ways in which a user can specify which friends have access to a piece of content on Facebook today. In the rest of the paper, we will use the term *social access control lists* (social ACLs or SACLs) to refer to such privacy policy specifications. A more precise definition is below.

Social access control lists (SACLs): A SACL is a privacy policy specification attached to a piece of content containing a proper subset of the user's friends; friends specified in the SACL have access to view and perform other actions on the content (e.g, liking or commenting). SACLs can be specified using different mechanisms provided by Facebook: allowing or denying access to individual friends one by one, specifying friend lists, using Facebook lists, or using a combination of handpicked friends and lists.[3]

It is important to note that SACLs *only* encompass custom settings by users and do not include the Facebook predefined access permissions: "everyone" or "public" (share with all Facebook users), "regional network" (share with

[2]When creating a SACL, a user can specify either individual friends or pre-created lists of friends; we refer to both of these as *terms*.

[3]Facebook also allows users who are *tagged* in a specific post to see the content [22, 24]. However, since users did not specify tagged friends explicitly through the privacy management interface, we do not consider them to be the part of SACLs. We leave exploring privacy expressed through tags to future work.

everyone in a regional network, deprecated in 2009), "all friends-of-friends" (share with all friends-of-friends), "all friends" (share with all friends), and "only me" (only visible to the user who uploaded the content).

2.2 Related work

Now that we understand the background of this paper, we discuss related work along three directions.

Understanding privacy awareness of users Researchers have studied the privacy awareness of social networking users [11, 30, 46]. These studies examine the profile information sharing behavior of users over a long period of time (e.g., 7 years) to understand if users' attitude towards their data privacy changes over time. Dey et al. [11] and Stutzman et al. [46] have shown quantitative evidence that Facebook users are sharing fewer profile attributes (such as hometown, birthdate and contact information) publicly over time. Social media discussions about Facebook privacy [8,13,44,45,47] and Facebook regularly rolling out new fine grained privacy management features [21] for the last few years have caught users' attention and potentially increased their concerns about available privacy controls.

How effective are users in managing their privacy settings? Recent studies have explored how effective users are in managing their privacy settings. Studies have shown that there exists a mismatch between desired and actual privacy settings when users share content on Facebook [4,5,29,34,36]. Liu et al. [34] conducted a user survey about privacy preferences for photos uploaded by users on Facebook. They found that privacy settings match users' expectations only 37% of the time, and when wrong, users are exposing their content to a much wider audience (e.g., all friends, friends-of-friends or even everyone on Facebook) than they intended. While the exact reason for incorrect privacy settings is hard to infer, it could be due to poor privacy management user interfaces or the significant cognitive burden required to manage privacy of their sensitive content.

Techniques for better privacy management Several techniques have been proposed to reduce the burden on users when managing their privacy settings. We can organize work in this space into two high level categories: (1) The first approach is to assist in automatically pre-defining grouping of friends that might be meaningful to the user for sharing sensitive content later. Facebook allows the user to pre-define such friend groupings using the friend list feature. But friend lists on Facebook have to be manually specified by the user today and this user overhead could be reduced by these approaches. (2) The second approach is to help the user on the fly to specify SACLs while sharing content. They predict SACL specifications with some input from the user. For example, if the user gives the name of a few friends that he wants to share content with, these approaches can automatically predict the remaining members of the SACL or provide recommendations of other possible SACL members. Now we will explain different proposals that fall in the above two categories.

PViz [39] is a proposal from first category that can automatically detect friend lists for a Facebook user and use it for better privacy policy visualization for the user. It leverages the network structure of the subgraph induced by the user's friends (i.e., the user's "one-hop subgraph") to detect friend lists using a modularity-based community detection algorithm. Using information extracted from friends' profile, it can also automatically assign a label to each detected list. This helps the user to understand the composition of a list. Based on these predefined lists, PViz points out to users which of her friends from a list can view a particular content. PViz presents a user study based on 20 users, who find PViz useful for understanding their existing privacy settings better.

However, many previous works [1, 16, 48] fall in the second category. They focus on recommending friends on the fly to the users as the user starts sharing a piece of content and selects a few intended friends. Privacy Wizard [16] is one example of such a tool. Privacy Wizard leverages network structure and profile attributes (like gender, age, education, work, etc) to recommend friends for inclusion in a privacy setting. The process starts as the user tags a few of her friends as "allowed" or "denied" for a content. It then uses a machine learning algorithm to classify the remaining untagged friends into an allow or denied category. The authors designed a survey experiment with 45 Facebook users and 64 profile data items to evaluate the accuracy of their tool. They observed that on average if a user tags 25 of her friends, the wizard configures her privacy setting with high accuracy. However in their experiments they did not look into the ground truth data on how a user actually specifies SACLs while sharing sensitive content.

We conduct a large scale study comprising of 1,165 users and all their uploaded content to focus on the privacy settings used by users. Most prior work tries to approximate ground truth privacy preference data by asking user privacy preferences explicitly, most of the time via surveys or via a combination of surveys and profile data collection using apps [30]. However, none of these studies looked into the "ground truth" data on SACLs (i.e., how a normal user would share their content using SACLs without any external intervention). To the best of our knowledge, we are the first ones to look into ground-truth data on users' usage of SACLs to propose insights on how to assist users to reduce the overhead of specifying SACLs.

2.3 Key questions

While fine-grained privacy controls put users in better control of their data privacy, it is not clear how users are using these privacy mechanisms. In this work, we take a first look at SACLs specified by 790 Facebook users (users who created at least one SACL) for 212,753 pieces of uploaded content. Our analysis focuses on the following key questions:

- *How are users using SACLs today?* We analyze how often SACLs are specified by users and how different types of content are shared using SACLs.
- *Is SACL membership predictable?* We analyze characteristics of SACL members to understand if they have something in common that other friends of the user do not. Our analysis explores whether members have similar profile attributes, exhibit strong social network connectivity with each other, or share similar activity levels. If we are able to separate out SACL members from among all friends, we may be able to automatically create SACLs for users.
- *What is the user overhead in specifying SACLs?* We

examine the overhead that users spend specifying SACLs today. The intuition is that, the more work required to create SACLs, the less usable the privacy mechanisms are.

- *What is the potential for reducing user overhead?* Based on our insights gained from analyzing SACL membership, we quantify the potential for further reducing the complexity of SACL specifications. Our findings serve as guidelines for designing better privacy management tools in the future.

3. DATASET

Now we describe the dataset we have collected about in-use SACLs on OSNs today.

3.1 Collecting data about SACLs in Facebook

Obtaining data at scale about user privacy specifications is quite challenging. In Section 2.2, we observed that most previous work used small-scale data about user privacy preferences. There are two main challenges in collecting large-scale data: First, we need permission to view the user SACLs. This is challenging as private settings on Facebook are private to the user; they cannot be obtained by crawling publicly visible user profiles. To address this challenge, we use the Facebook API [17], which offers methods to collect data about in-use privacy settings (provided the user gives us permission to do so). We therefore developed a Facebook application that helps users to better manage their privacy settings, and recruited users for the application. The Facebook application requests consent from the user to collect data about their SACL specifications for our research study. The data collection was performed under an approved Northeastern University Institutional Review Board protocol.

The second challenge is recruiting large number of users for the study who can provide consent to access their private SACL information. The traditional approach for recruiting users rely on personal communication (e.g., via email) or through an open call posted on a public bulletin board at a university or research lab. In such cases, the number of users that could potentially be recruited is usually limited to a few tens or hundreds. Another approach is to use online crowdsourcing platforms like Amazon Mechanical Turk (AMT) [38]. However, using a Facebook application violates the AMT policy that workers should not be required to download and install an application [2].[4] Instead, to recruit a variety of users at scale, we leverage the Facebook advertising platform.

3.1.1 Facebook Application: Friendlist Manager

We developed the Facebook app "Friendlist Manager" (or FLM) [25,35] that helps the user to automatically create and update friend lists that could be used for specifying SACLs. This application reduces the user burden of manually creating friend lists. FLM automates list creation by leveraging the network structure in the "one-hop subgraph" of the user. It uses network community detection algorithms [6, 40] to find overlapping communities in the one-hop subgraph. We found that users found FLM to be helpful; 480 (41%) of our users allowed FLM to update at least one of their lists. It is important to note that for the analysis in this paper, we only consider the content users shared and members of friend lists users had *before* installing FLM; this ensures that usage of our app does not impact the results.

When installing FLM, we request permission to access the following types of user data: basic user profile details including workplace, education, current city and family; privacy settings (including SACLs) used for all uploaded content (photos, videos, statuses, notes, music, questions, Shockwave Flash Player (SWF) movies, and checkins); and the friends, friend lists, and Facebook lists of the user. Should the user choose to not grant us access to their content, they are still allowed to use the application.

3.1.2 Recruiting users

To recruit users, we set up an advertisement for FLM on the Facebook advertising platform. The Facebook advertising platform allows us to reach out to the large Facebook population and target users with specific demographics. Our ad included the following text:

> *Need help to better organize your list of friends? Give FLM a try!*

Starting from June 20th 2012, we ran five ad campaigns for 10 days targeting 10 countries where English is an official language: USA, UK, Australia, New Zealand, Canada, India, Pakistan, Singapore, South Africa, and Philippines. In total, 232 users installed our app during this period. After this initial push, our app received a steady stream of new users through August 2013; in total, we observed 1,007 additional users after the advertising campaign ended. We believe FLM also spread "virally", with users "liking" or recommending the app to their friends. While it is hard to trace the source of these new 1,007 users, we found that 59 of them were friends of users who installed FLM through ads. The remaining users likely found FLM through search tools (e.g., Google Search) or through word-of-mouth based propagation.

Overall, a total of 1,239 users installed our application. For the purpose of this study, we only focus on the 1,165 users (94%) who gave us permission to access all the data we required for our research study.

3.1.3 User bias

One potential issue with user studies is a bias in the user population. In our case, it is challenging to obtain a random set of Facebook users. This is a fundamental issue with most user studies, and the common methodology is to carefully characterize the users under study to understand how diverse the users are in terms of demographics. Our user population is by no means random, and we report the demographics and behavior of users below. We believe that the users who installed FLM are those who are interested in creating friend lists to better manage their privacy settings. It is known that Facebook has been promoting friend lists as a way to more efficiently specify privacy preferences [14]. Thus, our user sample is most likely biased towards privacy-aware Facebook users. Additionally, ads can only be shown to users when they are logged in to Facebook; we are therefore likely to get users who are active on Facebook.

3.1.4 Ethical Considerations

The data we collected in FLM is highly privacy sensitive, and we took great care to respect users' privacy. First, we

[4] Our data collection methodology requires users to install a Facebook application.

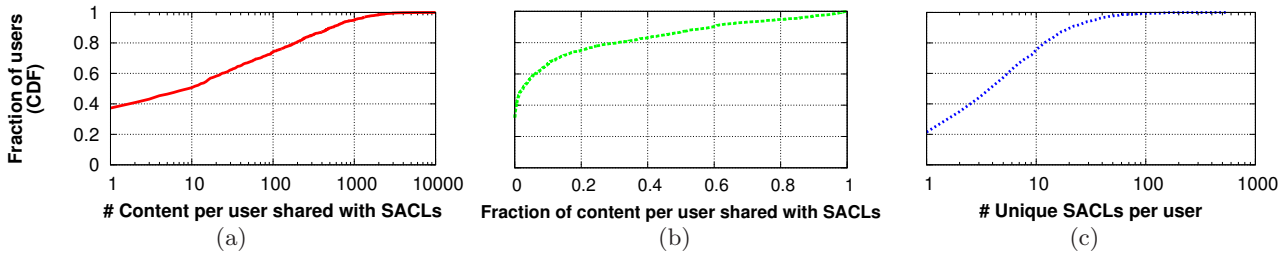


Figure 1: (a) Cumulative distribution of number of pieces of content uploaded by users using SACLs. A significant fraction (67.8%) of users in our dataset upload at least one content using SACLs. (b) Cumulative distribution of percentage of content per user shared using SACLs. More than 200 users in our dataset uploaded more than 30% of their content using SACLs. (c) Cumulative distribution of number of unique SACLs specified by each user who upload at least one content using SACLs.

conducted our study under Institutional Review Board approval. Second, we only report aggregate statistics here, and in any future papers. Third, we will never release any non-aggregated data to third parties. All of these steps were also included as part of FLM's Privacy Policy (provided to users when installing FLM).

3.2 FLM user demographics

We now examine the demographics of users who installed FLM and allowed us to collect their data. Users usually self-report their location, age, gender and education on their profile. We examine the "current location" attribute to estimate the location of users at a country level and find that 952 (82%) users have provided location information. According to this information, users are from 75 countries covering six continents. There were 19.2% users from North America, 18.1% from Europe and 35.5% from Asia. The top five countries were United States (20%), Pakistan (14%), India (7%), Brazil (7%), and Philippines (6%). Thus, we have users from a diverse set of geographic locations.

Next, we examine the age of users. Our users ranged between 18 and 65 and older, with the median age being 29; this distribution is in-line with the overall U.S. Facebook population [41]. Only 1.2% of users did not specify their gender and for the rest, we observed a strong male bias with 76% of our users being male; this differs from the overall U.S. Facebook breakdown of 47% male [41]. Finally, for the education level (reported by 73.8% of users), 67.8% of users have been to college, while 5.9% of them have been to graduate school. All these statistics demonstrate that we have a diverse set of users in our dataset.

As our users are recruited from a social network, one additional concern is that the users might be a "close-knit" group of friends (in terms of friendship), and not a more general sample of the user population. To evaluate whether this is the case, we check how closely related our users are by examining the number of users who are friends on Facebook, and the number of user pairs with at least one common friend. Out of the 678,030 possible pairs of users $[\binom{1,165}{2}]$, 44 (0.01%) were direct friends and 1,266 (0.19%) were not direct friends but had at least one friend in common. Thus, while our population does show some correlation with the social network (unsurprising, given the viral spreading we observed before), the user population is not strongly biased towards one small region of the entire Facebook social network.

Finally, we examine the activity of users in terms of uploaded content. Overall, our 1,165 users have an average of 518 friends (median 332), and uploaded on average 1,040 pieces of content (median 506). 1,003 (84%) users have uploaded more than 100 pieces of content. Only 39 (3.3%) users have uploaded fewer than 10 pieces of content while 3 (0.2%) of them have uploaded none. When we look at activity of users over time, we observe that the activities of our users spanned over 8 years from 2005 to 2013. We further find that 90% of users have been active for more than 20% of weeks since they joined Facebook. Our analysis of users suggests that we have a fairly diverse population most of whom are actively uploading content on Facebook.

4. SACL USAGE

We begin by examining the usage of SACLs by OSN users. Specifically, we investigate how often and for what types of content users specify SACLs.

1. How widely are SACLs used? We first examine how often different users share content with SACLs, using the FLM user set described in the previous section. Figure 1(a) presents the cumulative distribution (CDF) of the number of content shared using SACLs per user. We observe that a majority of our users are using SACLs for content sharing: 790 (67.8%) users out of 1,165 shared at least one piece of content using a SACL. In total, these 790 users uploaded 212,753 pieces of content using SACLs; this content accounts for 17.6% of all content uploaded by all 1,165 users. In the remainder of the paper, we focus only on these 790 users and the content they uploaded using SACLs. We note that the fraction of users using SACLs in our dataset is comparable to that reported for Google+ [31], where 74.8% of the users used SACLs. However, these Google+ users shared significantly more (67.8%) of their content with SACLs. This difference in the percentage of shared contents in Facebook and Google+ is likely due to the differences in user interface between the platforms. We leave a full exploration of the comparative use of SACLs across online social networks to future work.

Next, we observe that users use SACLs to different extents. In particular, we examine the percentage of content that each user shares with SACLs in Figure 1(b) (i.e., for each user, what fraction of their content is shared with a SACL?). We observe a biased distribution across users, but a significant fraction of users select SACLs for much of their

Content type	Total content items	Number shared with SACLs	Percent shared with SACLs
Status	786,800	139,112	17.7%
Photo	264,714	45,308	17.1%
Video	111,676	20,880	18.7%
Album	26,527	4,415	16.6%
SWF	9,794	1,554	15.9%
Note	8,500	883	10.4%
Checkin	3,224	548	17.0%
Question	374	25	6.7%
Music	355	27	7.6%
Offer	9	1	11.1%
Total	1,211,973	212,753	17.6%

Table 1: Distribution of the number and percentage of content shared with SACLs across different types of content.

content: 20% of users share more than 30% of all their content using SACLs.

Thus, we observe that SACLs are widely used by our users for sharing content, which encourages us to further explore the composition of SACLs and complexity of SACL specification in the following sections.

2. How many SACLs do users need to create? Having observed that SACLs are widely used, we now investigate how many different SACLs users create from amongst their friends. Figure 1(c) shows the cumulative distribution of the number of unique SACLs specified by each user. A large fraction (75%) of the users use less than 10 SACLs, and 20% of the users use only a single SACL. However, there are 5 heavy SACL users, who have used more than 100 unique SACLs. We find that these are heavy users of privacy settings and use different combination of a small number of lists and a set of handpicked friends to specify multiple SACLs for multiple pieces of content. Overall, most users only require a limited number of SACLs to share sensitive content; we leverage this finding later in Section 6 to reduce the user overhead in specifying SACLs.

3. Does SACL usage vary with content types? Facebook allows users to upload a variety of content types. Table 1 presents a breakdown of the total number of content items of different types, and the fraction of those items shared with SACLs. We are interested in understanding if users are biased towards a few types of content when using SACLs. The third and fourth columns of Table 1 show the number and fraction of each type of content shared using SACLs. We observe that SACLs are used across all nine different types of content. In fact, 10-20% of almost all types of content are shared with SACLs. Questions and music are shared least often with SACLs; we suspect that these types of content tend to be more public and are usually not privacy sensitive. This widespread use of SACLs across all types of content further justifies looking deeper into SACL membership and complexity, with the goal of increasing the usability of SACLs.

4. How are SACLs created? Facebook users can construct their SACLs in different ways. As mentioned in Section 2.1, while creating a SACL the user may allow or deny access to individual friends, or lists, or use a combination of friends and lists. Table 2 shows the distribution of number

	SACL created using only lists	only friends	both lists and friends
Number of Users	593	555	213
Percent of users using SACLs	75.9%	71.1%	27.3%
Percent of SACLs	33.5%	44.3%	22.2%
Percent of content shared with SACLs	61.4%	27%	11.6%

Table 2: Distribution of the number of users using different mechanisms to create SACLs while sharing contents.

of users using different mechanisms to create SACLs. We observe that more than 70% of users are creating at least one SACL by individually selecting their friends and more than 44% of SACLs fall in this category; this is surprising, as selecting friends individually is a somewhat tedious task. Interestingly, only 27% of SACL content is shared with such SACLs; users share the majority of their content with SACLs created using lists.

5. How many users are in SACLs? Next, we examine the size of SACLs (i.e., how many of a user's friends are in different SACLs). Figure 2 presents a CDF of fraction of the SACL owner's friends that the SACLs contain. We observe that the distribution exhibits three distinct regions, described below:

1. **Include only few friends:** The first region is highlighted in gray on the left side of the graph; this region contains SACLs with between 0% and 5% of the user's friends. This region contains 25% of all SACLs. In the remainder of the paper, we refer to these as include few SACLs.

2. **Exclude only few friends:** The next region is highlighted in gray on the right side of the graph; this region contains SACLs with between 95% and 100% of the user's friends. This region contains 26% of all SACLs. In the remainder of the paper, we refer to these as exclude few SACLs.

3. **Include subset of friends:** The final region is in the middle of the graph; this region contains between 5%

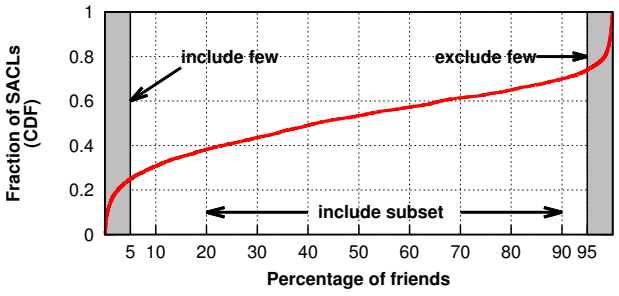


Figure 2: Percentage of friends included in SACLs. SACLs in the left gray region are the include few SACLs, SACLs in the right gray region are exclude few SACLs, and the SACLs in middle white region are the include subset SACLs.

and 95% of the user's friends. This region contains the plurality (49%) of the SACLs. In the remainder of the paper, we refer to these as include subset SACLs.

As we suggest below, the distribution of SACL sizes is very likely influenced by the interface for SACL specification. We use our categorization in the rest of the chapter when we try to characterize the SACL members across different features. However, for exclude few SACLs we also want to see whether we can characterize the excluded friends; for these, we also examine the *excluded members of exclude few SACLs*. The plurality of include subset SACLs shows that our users are not simply including or excluding a handful of their friends, but are often including large subsets of their friends. This result further motivates the need to understand how these subsets are selected.

Overall, our observations suggest SACLs are being widely used today by a majority of our users to control access to a non-trivial fraction of their content. SACLs are used at different rates by different users, but they do appear to be used to share many different types of content. Finally, SACLs show wildly different sizes, with many SACLs containing few or almost all of a user's friends. With this understanding, we turn to examine the membership of SACLs in the following section.

5. SACL MEMBERSHIP

We now take a closer look at the membership of SACLs. In other words, are the members of a given SACL distinguishable from the SACL creator's other friends? (e.g., do the members share a profile attribute?) This question is interesting to examine, as any automatic detection of SACL membership would only work if the SACL members were distinguishable. Moreover, existing work [16, 39, 48] hypothesizes that profile attributes, network structure, and user activity can help us to automatically detect clusters corresponding to SACLs; we aim to see if this is true using our dataset of real-world fine-grained privacy settings.

5.1 Methodology

Our analysis in this section explores the possibility of characterizing SACL members as a group across three features: (i) profile attributes, (ii) social network structure, and (iii) activity. In other words, we would like to see whether the SACL members form a distinct cluster among the friends of the user. To do so, we form clusters based on these three features and then examine how closely the SACLs of the user match our cluster (e.g., we form a cluster of all user's friends who attended the same high school and then look to see if this cluster matches any SACL). To compare our automatically detected clusters and the user's SACLs, we address three separate questions:

1. Do the automatically detected clusters match SACLs? Once we have the clusters of friends for a given feature, for each SACL, we try to find the best matching cluster. To compute the "goodness" of a match, we use the F-score metric [37] which provides a measure of detection accuracy. It is computed as the harmonic mean of precision and recall; F-score varies from 0 to 1, with 1 representing a perfect match.

Unfortunately, a low F-score does not necessarily imply that SACLs are not correlated with automatically detected groups. For example, the members of a SACL could be split between two automatically detected groups; in this case, the F-score for both groups would be quite low, but the F-score of the union of the groups would be quite high. Looking deeper into this issue takes us to our next question.

2. How distributed are the SACLs across clusters? In order to check how widely the SACL members are spread across the clusters we use the metric *entropy* [10]. For a given SACL and a cluster c from a set of automatically detected clusters C, we can compute $p(c)$, the probability of a SACL member belonging to c. Then we measure the *entropy* of the SACL as

$$-\sum_{c \in C} p(c) \log_2 p(c)$$

A higher value of entropy signifies more diversity within the SACL members (i.e., they are spread across more clusters).

To be able to compare across the SACLs which belong to different users (with different numbers of clusters and friends per cluster), we normalize the entropy using maximum possible entropy per SACL. A SACL will have maximum entropy when its members are uniformly distributed across all clusters [10]; in this case the entropy will be $\log_2 |C|$, where $|C|$ is the number of clusters in C. We therefore calculate *normalized entropy* as

$$-\frac{\sum_{c \in C} p(c) \log_2 p(c)}{\log_2 |C|}$$

Normalized entropy for a SACL ranges from 0 to 1. A normalized entropy close to 1 indicates that a SACL is uniformly spread across the maximum number of clusters and a normalized entropy close to 0 indicates that all or most of the SACL members are part of one cluster.

3. How are SACLs different from random groups? One outstanding issue remains: We are examining the entropy of SACLs, but we would really like to measure what's the likelihood of selecting the members of a SACL by pure chance. For example, suppose all of a user's friends attended the same high school; in this case, the "high school" cluster (all friends in a single cluster) would perfectly match any large SACL.

To measure the uniqueness of SACLs relative to random groups, we use the Adjusted Rand Index (ARI) [43] to determine the similarity between SACL and automatically detected clusters. ARI is a similarity metric normalized against chance and varies from -1 to 1. An ARI of 0 indicates no better similarity than a random group, a negative ARI implies worse similarity than a random group, and an ARI of 1 indicates exact similarity. For each SACL, we calculate at the ARI provided by the most similar cluster. If most of the SACLs have ARI close to 0, then the automatically detected clusters are no better in detecting the SACLs than simply using random groups.

5.2 SACLs and user attributes

We first explore whether SACL membership is correlated with a common profile attribute. To do so, we leverage the profile attributes provided by Facebook users, focusing on four attributes: workplace, education, current city, and family. We choose these attributes as they have been shown be most strongly correlated with groupings of users in social networks [40]. Using these attributes, we group the friends

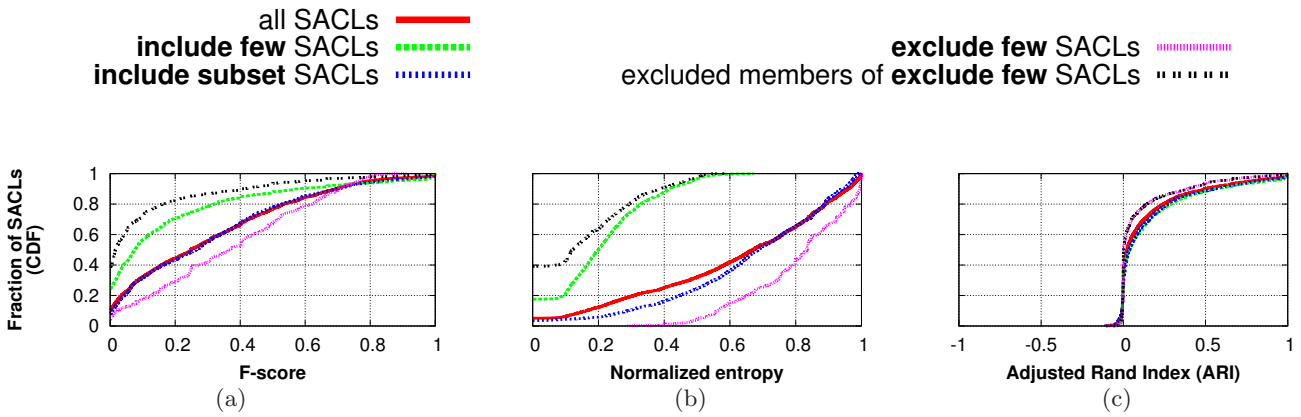


Figure 3: Correspondence between the attribute-based clusters and SACLs, with the cumulative distributions of (a) F-score, (b) Normalized entropy, and (c) ARI. Figure 3(a) shows only 15% of the automatically generated attribute-based communities have a F-score of more than 0.6, indicating low number of SACLs showing high match. Figure 3(b) shows that larger SACLs are spread across multiple such clusters and have higher normalized entropy. The reverse is true for include few SACLs and excluded members of exclude few SACLs. However, Figure 3(c) confirms that more than 40% of these SACLs show better similarity with attribute-based clusters than random groups.

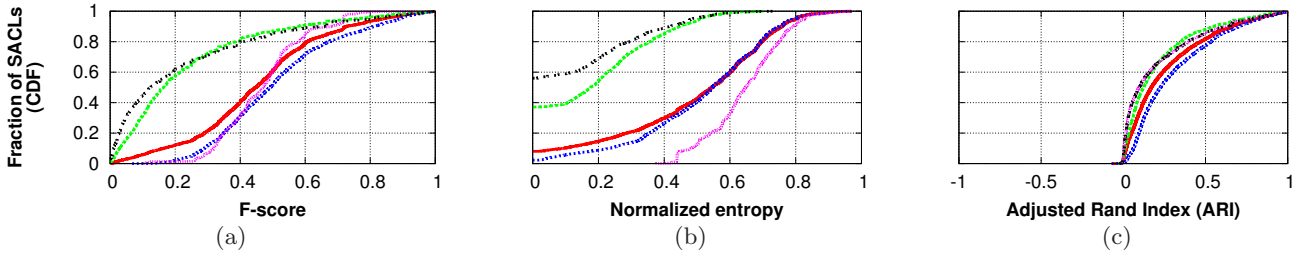


Figure 4: Correspondence between the network communities and SACLs. Figure 4(a) shows 21% of network communities have a F-score of more than 0.6, indicating a relatively poor match between network communities and SACLs. Figure 4(b) and Figure 4(c) confirms that though the larger SACLs have higher entropy (i.e., they are distributed across multiple communities), more than 90% of these SACLs show better similarity with network-based clusters compare to random groups.

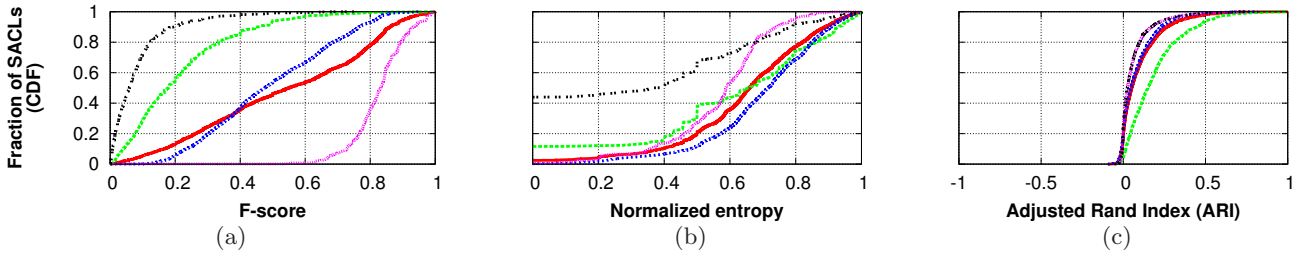


Figure 5: Correspondence between the activity-based clusters and SACLs. Figure 5(a) shows 47% of the automatic attribute clusters shows a F-score of more than 0.6, indicating comparatively strong match between activity communities and SACLs. However, Figure 5(b) shows that the larger SACLs have higher entropy, and Figure 5(c) shows that only 4% have ARI more than 0.3. As a result, random groupings of friends the same size as SACLs would likely show a degree of similar matching.

of a user into clusters who share a common attribute. We report results for all attribute groups in aggregate for brevity; the results are similar when considering each attribute type alone.

We begin by using the F-score metric to check how many of the SACLs exactly match the attribute-based clusters. We present the cumulative distribution of F-scores across all SACLs in Figure 3(a). The figure shows that only 15% of all SACLs have a F-score of more than 0.6, indicating a good match for a small subset of SACLs. The result is even worse for very small SACLs (include few or the excluded members of exclude few), with only 10% of such SACLs having a F-score more than 0.6.

We explore the reason for the low F-scores by analyzing the normalized entropy of these SACLs in Figure 3(b). The figure shows that the small SACLs have a low entropy (with 20% of include few SACLs with entropy 0) indicating they are mostly part of single attribute-based clusters (this is unsurprising, given that these SACLs are small). On the other hand, the larger SACLs show a high entropy with 35% of exclude few SACLs having an entropy of more than 0.8. These results suggest that our attribute clusters are overestimating the smaller SACLs (indicated by low entropy and low F-score) and underestimating the larger SACLs (indicated by high entropy and low F-score).

Finally, we examine whether SACLs match attribute clusters better than random groups using ARI. As mentioned in Section 5.1, an ARI of 0 indicates similarity no better than random groups. Figure 3(c) presents the cumulative distribution of ARI across all SACLs; we observe that 68% of all SACLs have ARI larger than 0, indicating they have more similarity with attribute-based clusters than a purely random set of friends.

Overall, our results suggest that only a small number of attribute clusters serve as a close match for SACLs. However, the SACLs do show some correlation with attribute groups when compared to random subsets of the user's friends. Next, we look into the correlation between SACLs and the social network to see if network-based clusters more closely approximate the SACLs.

5.3 SACLs and network structure

In order to explore whether the SACLs correspond to the network structure, we first identity clusters of the user's friends that are tightly connected in the social network (these clusters are often called *network communities*). Specifically, we extract all of the friendship connections between the user's friends, and then use a community detection algorithm that has been shown to work well in grouping a user's friends into a small set of clusters [35]. This algorithm is a combination of a global community detection algorithm [6] and a local community detection algorithm [40] to detect overlapping communities.[5]

We begin by examining how many of the SACLs exactly match one of the social network-based clusters. Figure 4(a) presents the cumulative distribution of F-scores across all SACLs. The figure shows that 21% of all SACLs have a F-score of more than 0.6, indicating a good match for 21% of SACLs. This is significantly higher than the attribute-based clusters in the prior section, but still does not show a strong correlation.

We next analyze the normalized entropy of these SACLs in Figure 4(b). Similar to the attribute-based clusters, the network communities tend to overestimate smaller SACLs and underestimate larger SACLs, but at a much lower rate. We verify these findings using ARI in Figure 4(c). We can observe that 98% of all SACLs have ARI more than 0, indicating almost all of the SACLs have more similarity with community based clusters than a purely random set of friends.

Overall, the network communities show better match with SACLs compared to the attribute clusters. However, still only a small fraction of SACLs have strong correlation with network communities, making it unlikely that network communications could be used to infer SACLs for sharing content. Next we will look into the correlation between activity-based clusters and SACLs.

5.4 SACLs and user activity

For our final user feature, we examine whether the membership of SACLs is correlated with the strength of the link between the user and their friends. As a proxy for link strength, we use *activity*; this is a common way to estimate how closely connected two users are [3]. For each user, we collected data about four different types of interaction between the user and their friends: (i) posting on the user's wall, (ii) liking the user's posts, (iii) commenting on the user's posts, and (iv) being tagged in the user's status and photos. We observe that 94% of the users who used SACLs have at least one such interaction with their friends.

Using this data, we cluster each user's friends by activity (i.e., frequency of interaction) and see whether the activity-based clusters matched the SACLs. We use the same algorithm as prior work [3] to find the activity clusters. The algorithm is essentially a k-means algorithm modified to automatically find the optimal number of clusters. As a result, all friends with a similar number of interactions will be put in one cluster. After running the algorithm, we find that the median number of clusters across all users is four.[6]

As before, we begin by examining the cumulative distribution of the F-score in Figure 5(a); we observe that 47% all SACLs have a F-score of greater than 0.6. This is even better than network-based communities. The larger SACLs (e.g., exclude few SACLs) show an even stronger match with high F-scores, but the match is considerably worse for smaller SACLs (e.g., include few SACLs), with only 3% of such SACLs having a F-score more than 0.6. Closely examining the activity-based clusters, we hypothesize that our method of creating activity communities often results in creating a large cluster containing all friends with low levels of interaction with the user. As a result, this single, large community alone overlaps with the large SACLs considerably, making their F-score quite high.

To confirm this hypothesis, we also calculate the cumulative distribution of normalized entropy (Figure 5(b)) and ARI (Figure 5(c)). We find a poor match between SACLs

[5] We note that there are a large variety of community detection techniques in the literature. To make sure our choice of algorithm did not bias the results of our analysis, we performed the same analysis with two additional community detection algorithms [9, 42] similar to ones used in earlier work on unsupervised detection of privacy settings [16, 39]. Our results were similar with these algorithms, and so we omit the results for brevity.

[6] Interestingly, this observation matches Dunbar's sociological study [12, 28] where the number of Dunbar's circles, the number of activity-based clusters in people's offline network is also four.

Figure 6: Facebook's interface for specifying SACLs.

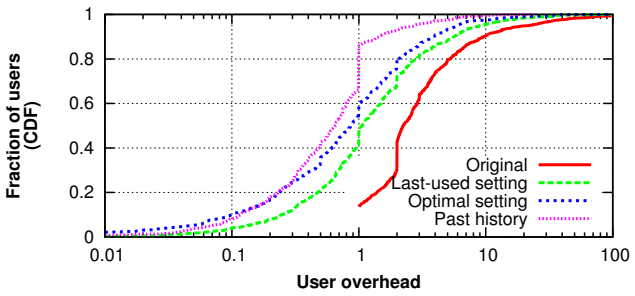


Figure 7: Cumulative distribution of overhead for specifying SACLs. Shown are the distributions of measured SACLs (Original), measured SACLs taking into account Facebook's last-used setting (Last-used setting), the optimal overhead for measured SACLs (Optimal), and the overhead of our proposed mechanism of presenting the user with the last 5 SACLs (Past history). Our proposed mechanism shows a substantial reduction in user overhead.

and activity-based clusters using both pieces of analysis; the ARI values for the SACLs are very close to 0 for almost all activity-based clusters (e.g., only 8% of the SACLs have ARI more than 0.3). This finding confirms that any random groups of the size of larger SACLs will show the same level of similarity with the activity-based clusters, thus making the clusters a poor mechanism for approximating SACL membership.

5.5 Summary

In this section, we examined the membership of SACLs by trying to correlate SACL members with attribute, network structure, and activity-based clusters. Our results show that very few of these clusters show a significant correlation with SACLs, suggesting that automatically detected SACLs-based on these features are unlikely to be very accurate. This finding is in opposition to the results from prior work [16, 39, 48], which suggest that it is possible to use automatically detected clusters to create SACLs. We believe this difference is due to the fact that these prior works were not able to evaluate their proposals against ground-truth SACLs. In fact, others have also found [33] that users are able to group their friends in meaningful groups, but find it difficult to choose the right group to share content with. Consequently, we explore alternative approaches to increase the usability of SACLs in the next section.

6. SACL SPECIFICATION

We have observed that SACLs appear to be quite difficult to infer automatically. We now examine the "overhead" (i.e., the amount of work that users must perform) in order to specify SACLs today. Then, we explore how we can increase the usability of these SACLs by reducing the user overhead and making SACLs easier to use. If successful, these approaches would make privacy easier for users to manage, thereby increasing the usability of OSNs in general.

6.1 SACL specification overhead

The act of specifying a SACL—choosing which friends to share content with—induces cognitive overhead on the user. While there may be multiple dimensions of this overhead (e.g., deciding whether to include a specific user, using the interface, etc), many of these are quite challenging to measure. As a first step, we define the *SACL specification overhead* to be the number of *terms* used to specify a SACL. Our reasons for doing so is that Facebook allows users to specify SACLs using an allow/deny interface, where users can select friends or lists to allow or deny access (a screenshot of Facebook's interface is shown in Figure 6). Thus, the amount of work the user has to do is proportional to the number of friends/lists that the user selects to allow or deny. Of course, we recognize there are dimensions of overhead that this measure fails to capture; we leave the task of characterizing those dimensions of user overhead to future work.

As an example, consider the screenshot shown in Figure 6. In this example, the user is choosing to allow friends Bob, Carol, the list Football Friends, and the Facebook list Family. The user is also choosing to deny the list Drinking Buddies. As a result, the SACL specification overhead for this SACL is five (A total of five terms appear in the allow and deny settings.) It is important to note that the size of the SACL is different from its specification overhead: Consider the case of a user only denying access to a single friend. In this case, the specification overhead is low, but the SACL has many users in it.

We define the *average user overhead* as the average of all SACL specification overheads for content shared by a given user. Formally, if a user specifies access to her content $\{c_1, c_2, ...c_n\}$ with privacy settings $\{p_1, p_2, ...p_n\}$, the average user overhead for this user is

$$\frac{\sum_i |p_i|}{n}$$

where $|p_i|$ denotes the SACL specification overhead of setting p_i. Note that the p_i settings are not necessarily distinct, as multiple pieces of content may be shared with the same SACL. The best-possible average user overhead is 1, meaning the user only used the SACLs with a single term when sharing her content.

We present cumulative distribution of the average user overhead in Figure 7 with the line "Original". We ob-

serve that users do have significant overhead when specifying SACLs: 86% of users have an overhead more than 1, and there are more than 150 users with overhead more than 5. This suggests there is significant potential to reduce the SACL specification overhead for users, making SACLs more usable.

6.2 Last-used setting

Facebook's default privacy setting for content is set to select the *last-used* privacy setting [20]. So, if a user selects a SACL for a newly uploaded piece of content, all future pieces of content will be shared with the same SACL until the user chooses a different privacy setting. We therefore modify our definition of average user overhead to capture this behavior; if a user selects the same SACL as the previous piece of content, we define this SACL specification overhead to be 0. As a result, a user's average user overhead may be less than one.

We present the cumulative distribution of the average user overhead, taking into account the last-used setting, in Figure 7 with the line "Last-used setting". We immediately observe a significant reduction in the measured overhead, which we believe more accurately captures the work a user must do. The figure shows that this simple technique of using last setting as the default can significantly reduce the user overhead: this technique lowers the overhead by more than 50% for almost half (48%) of the users. Thus, Facebook's choice to enable default last-used settings is useful in reducing user overhead. For the remainder of our analysis, we use average user overhead, taking into account the last-used setting, as our baseline.

6.3 Optimal overhead

It is important to note that there may be multiple ways of specifying a given SACL: For example, a user could specify the SACL by only allowing the friends in the SACL. Or, the user could use an existing list, and exclude the users not allowed to access the content. Or, the user could allow all friends, and then deny only the friends who should not be able to access the content. We now examine how close the user's chosen specifications are to the *optimal specification*, in terms of the SACL specification with the minimum overhead.

To do so, for each SACL we observe, we determine the optimal specification overhead.[7] We then present the cumulative distribution of average user overhead in the optimal case in Figure 7 with the line "Optimal setting". We observe that there is still room for improvement from using the last-used setting alone; many users could express their SACLs in a manner than involves fewer terms.

6.4 Using past history

In this section, we explore a generalization of the last-used setting, with the goal of further reducing the average user overhead. The results in Section 4 suggested that there are certain SACLs that users select to share content with a significant fraction of the time. Figure 8 plots the cumulative distribution of the percentage of content shared with the top k most frequently used SACLs for each user. For example we can see that if we allow each user to use their top 5 SACLs, this would cover over 80% of the content for the vast majority (90%) of users.

[7]Note that computing the optimal overhead of a setting is a modified version of the NP-hard set cover problem [32] where the setting is the universe and lists and individual friends are subsets of the universe. We use a brute force solution to the problem, which is feasible due to small number of subsets in this case.

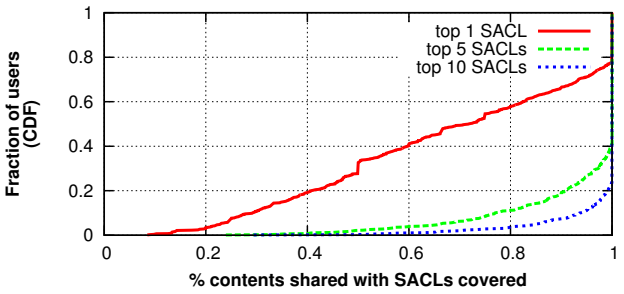


Figure 8: Cumulative distribution of percentage of content covered by the top k SACLs. Even if we set k=5, most of the content for the majority of users are covered.

This observation means that we may be able to significantly reduce the average user overhead by allowing users to choose from their k *most used* SACLs, rather than just the last-used SACL. To do so, we calculate the average user overhead, assuming one would have made it possible for the user to directly use the top 5 most frequently used settings while sharing content (Should the user re-use these settings, we calculate the overhead as 0). A cumulative distribution of the resulting overhead is shown in Figure 7 with the line "Past history". We immediately observe a dramatic reduction in user overhead (In fact, the overhead is lowered for 86% of users).

In summary, this approach of leveraging past history has the potential to significantly reduce the user overhead in specifying SACLs. An OSN operator can create these SACLs based on user's past history, and provide them as options to select from, when the user uploads a new piece of content. Should the user select one of the previously-used SACLs, it will reduce their overhead and make privacy specification more usable.

7. CONCLUDING DISCUSSION

Online social networks are increasingly popular and users are sharing ever more content on these services. In this paper, we focused on the most privacy-sensitive of these content: the content with hand-crafted privacy settings selected by the users. We found that these SACLs are surprisingly common (over 17.6% of all content is shared with SACLs), but that the membership of these SACLs shows relatively little correlation with the profile attributes, the social network structure, or the activity level of the members. As a result, there appears to be little hope of automatically detecting more than a few of these SACLs. We also found that the act of specifying SACLs is often complicated for users, but a simple technique like remembering a few of the most frequently used SACLs is likely to significantly reduce this burden in practice.

However, much work remains to be done. In the remainder of this section, we discuss a few of the limitations of our study, as well as future directions for exploration.

Understanding motivation for SACLs Our study explores the use of SACLs, but does not reveal *why* users create SACLs or how they choose the content to share with SACLs. Possible reasons include dissatisfaction with the default privacy settings, the sharing of highly privacy sensitive content, or using SACLs as a mechanism to choose the audience for a particular content.

Moving target We quantified the way users create in-use SACLs today, but Facebook is known for changing their privacy interface over time [18]; these changes are likely to impact the usage of SACLs for individual users. We aim to repeat our analysis as Facebook makes these changes, hoping to capture resulting changes in user behavior.

SACL accuracy It remains an unexplored question as to which of the friends users would *ideally* want to share their content with (i.e., who does the user want to be in a SACL, regardless of who is actually in the SACL). Prior work has shown that users often misunderstand other Facebook privacy settings [34], and we suspect that this would likely hold true for SACLs as well.

SACL overhead In our calculation of overhead, we took into consideration the number of terms specified by users explicitly while specifying SACLs, where a term can be a friend list or an individual friend. However, this quantification does not directly account for the mental effort required for a user when creating SACLs (e.g., certain SACLs may be easier or harder to create, even if they have the same number of terms). We leave a full exploration of this effect (possibly via detailed debriefing interviews of a small sample of Facebook users) to future work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. We also thank our FLM users for enabling us to conduct this study. This research was supported in part by NSF grants CNS-1054233 and CNS-1319019, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education Grant.

8. REFERENCES

[1] S. Amershi, J. Fogarty, and D. Weld. Regroup: interactive machine learning for on-demand group creation in social networks. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, Austin, TX, May 2012.

[2] Amazon Mechanical Turk terms of service. `https://requester.mturk.com/mturk/help?helpPage=policies`.

[3] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni. Analysis of Ego Network Structure in Online Social Networks. In *Proceedings of the 4th ASE/IEEE International Conference on Social Computing (SocialCom'12)*, Amsterdam, The Netherlands, September 2012.

[4] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the 31st SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*, Paris, France, April 2013.

[5] A. Besmer and H. R. Lipford. Moving Beyond Untagging: Photo Privacy in a Tagged World. In *Proceedings of the 28th SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, Atlanta, GA, April 2010.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of community hierarchies in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[7] D. Boyd and E. Hargittai. Facebook privacy settings: Who cares? *First Monday*, 15(8), 2010.

[8] S. Chen. Can Facebook get you fired? Playing it safe in the social media world. `http://www.cnn.com/2010/LIVING/11/10/facebook.fired.social.media.etiquette/index.html`, 2010.

[9] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:1–6, 2004.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[11] R. Dey, Z. Jelveh, and K. W. Ross. Facebook users have become much more private: A large-scale study. In *Proceedings of the 10th Annual IEEE International Conference on Pervasive Computing and Communications (perCom'12)*, Lugano, Switzerland, March 2012.

[12] R. I. M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 1998.

[13] A. Etzioni. Despite Facebook, privacy is far from dead. `http://www.cnn.com/2012/05/25/opinion/etzioni-facebook-privacy/`.

[14] Facebook smart lists. `https://blog.facebook.com/blog.php?post=10150278932602131`.

[15] Number of active users at Facebook over the years. `http://finance.yahoo.com/news/number-active-users-facebook-over-years--214600186--finance.html`.

[16] L. Fang and K. LeFevre. Privacy Wizards for Social Networking Sites. In *Proceedings of the 19th International World Wide Web Conference (WWW'10)*, Releigh, NC, April 2010.

[17] Facebook API. `http://developers.facebook.com/docs/reference/api/`, 2011.

[18] Detailed History of Facebook Changes 2004-12. `https://www.jonloomer.com/2012/05/06/history-of-facebook-changes/`.

[19] The Evolution of Privacy on Facebook – Changes in default profile settings over time. `http://mattmckeon.com/facebook-privacy/`, 2010.

[20] When I post something, how do I choose who can see it? `https://www.facebook.com/help/120939471321735`.

[21] New Tools to Control Your Experience. `https://blog.facebook.com/blog.php?post=196629387130`.

[22] Making Photo Tagging Easier. `https://blog.facebook.com/blog.php?topic_id=203150980352`.

[23] Facebook's New Privacy Changes: The Good, The Bad, and The Ugly. `https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly`.

[24] Tag Friends in Your Status and Posts. `https://www.facebook.com/notes/facebook/`

tag-friends-in-your-status-and-posts/109765592130.

[25] Friendlist Manager. http://friendlist-manager.mpi-sws.org/.

[26] More Privacy Options. https://blog.facebook.com/blog.php?post=11519877130.

[27] Facebook's New Groups, Dashboards, and Downloads Explained. http://www.fastcompany.com/1693443/facebooks-new-groups-dashboards-and-downloads-explained-video.

[28] R. A. Hill and R. I. M. Dunbar. Social network size in humans. *Human Nature*, 14(1), 2003.

[29] C. M. Hoadley, H. Xu, J. J. Lee, and M. B. Rosson. Privacy as information access and illusory control: The case of the Facebook News Feed privacy outcry . *Electronic Commerce Research and Applications*, 9(1), 2010.

[30] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and Privacy: It's Complicated. In *Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12)*, Washington, DC, July 2012.

[31] S. Kairam, M. Brzozowski, D. Huffaker, and E. Chi. Talking in circles: selective sharing in google+. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, Austin, TX, May 2012.

[32] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, 1972.

[33] P. G. Kelley, R. Brewer, Y. Mayer, L. F. Cranor, and N. Sadeh. An Investigation into Facebook Friend Grouping. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction (INTERACT'11)*, Lisbon, Portugal, September 2011.

[34] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the 11th ACM/USENIX Internet Measurement Conference (IMC'11)*, Berlin, Germany, November 2011.

[35] Y. Liu, M. Mondal, B. Viswanath, M. Mondal, K. P. Gummadi, and A. Mislove. Simplifying Friendlist Management. In *Proceedings of the 21st International World Wide Web Conference (WWW'12)*, Lyon, France, April 2012.

[36] M. Madejski, M. Johnson, and S. M. Bellovin. The Failure of Online Social Network Privacy Settings. Technical Report CUCS-010-11, Department of Computer Science, Columbia University, 2011.

[37] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[38] W. Mason and S. Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 2011.

[39] A. Mazzia, K. LeFevre, and E. Adar. The PViz comprehension tool for social network privacy settings. In *Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12)*, Washington, DC, July 2012.

[40] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in Online Social Networks. In *Proceedings of the 3rd ACM International Conference of Web Search and Data Mining (WSDM'10)*, New York, NY, February 2010.

[41] D. Noyes. The Top 20 Valuable Facebook Statistics. http://zephoria.com/social-media/top-15-valuable-facebook-statistics/.

[42] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2), 2004.

[43] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 1971.

[44] People Manage Their Privacy On Facebook Naturally. http://www.sciencedaily.com/releases/2009/04/090420084957.htm.

[45] More Facebook Friends Means More Stress, Says Report. http://www.sciencedaily.com/releases/2012/11/121126131218.htm.

[46] F. Stutzman, R. Gross, and A. Acquisti. Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. *Journal of Privacy and Confidentiality*, 4(2), 2012.

[47] J. D. Sutter. Some quitting Facebook as privacy concerns escalate. http://www.cnn.com/2010/TECH/05/13/facebook.delete.privacy/index.html.

[48] Q. Xiao, H. H. Aung, and K.-L. Tan. Towards ad-hoc circles in social networking sites. In *Proceedings of the 2Nd ACM SIGMOD Workshop on Databases and Social Networks(DBSocial'12)*, Scottsdale, AZ, May 2012.

To Befriend Or Not? A Model of Friend Request Acceptance on Facebook

Hootan Rashtian, Yazan Boshmaf, Pooya Jaferian, Konstantin Beznosov
University of British Columbia, Vancouver, Canada
{rhootan,boshmaf,pooya,beznosov}@ece.ubc.ca

ABSTRACT

Accepting friend requests from strangers in Facebook-like online social networks is known to be a risky behavior. Still, empirical evidence suggests that Facebook users often accept such requests with high rate. As a first step towards technology support of users in their decisions about friend requests, we investigate why users accept such requests. We conducted two studies of users' befriending behavior on Facebook. Based on 20 interviews with active Facebook users, we developed a friend request acceptance model that explains how various factors influence user acceptance behavior. To test and refine our model, we also conducted a confirmatory study with 397 participants using Amazon Mechanical Turk. We found that four factors significantly impact the receiver's decision, namely, knowing the requester's in real world, having common hobbies or interests, having mutual friends, and the closeness of mutual friends. Based on our findings, we offer design guidelines for improving the usability of the corresponding user interfaces.

1. INTRODUCTION

Users of Facebook-like online social networks (FOSN) are not careful when accepting friend requests from strangers, i.e., those who they do not know in real life or online communities [3, 20]. This behavior can be exploited by an attacker to run an infiltration campaign in a target FOSN [6]. Such malicious campaigns are a growing cyber-security threat [9], where an attacker controls a set of user accounts and exploits them to befriend a large number of benign users.

Large-scale infiltration has three alarming security implications [6]: First, the social graph of the target FOSN is compromised and polluted with a large number of non-genuine social relationships. This means that third-party services and websites have to perform appropriate "cleaning" to mask out fake accounts and their relationships before integrating with or using such a FOSN. Second, and other than online surveillance, the attacker can breach the privacy of users and collect large amounts of personally identifying information (PII), such as email addresses, phone numbers and birthdates, which have considerable monetary value in the Internet underground markets [5]. In addition, this information can be used to run follow up, highly personalized e-mail spam and phishing campaigns [16]. Third, the attacker can exploit the infiltrated FOSN to spread misinformation as a form of political astroturfing [23], or even influence algorithmic trading that uses opinions extracted from FOSNs to predict stock markets [2, 4].

Preventing large-scale infiltration, or at least limiting its scale and impact, is important not only to users but also to FOSN operators and social media-based businesses. Improved technology support for FOSN users in helping them to make better decisions in regards to friend requests is expected to reduce the associated risk. This, however, requires a better understanding of user's befriending behavior in FOSNs, particularly what makes them to accept or decline friendship requests.

Our research bridges this knowledge gap. In particular, we aim to answer the following general research question: *Why do FOSN users accept friend requests from strangers?* In our studies, we focused on the scenario where a FOSN user receives a friend request from another, a stranger in particular, and investigated the factors that influence the user's decision on whether to accept this request. Moreover, we also studied the process that users go through, when accepting friend requests, including identity verification, new friend management, and privacy settings updates.

In order to understand users' behavior in FOSNs, we designed two studies: a qualitative, exploratory study and a quantitative, confirmatory study. We received an approval for both studies from our university's research ethics board.

First, we conducted a set of semi-structured interviews with 20 active Facebook users (Section 2). The goal of conducting this exploratory study was to understand users' behavior in FOSNs in response to friend requests, and explore the factors that influence their decisions. To the best of our knowledge, there is no related qualitative work to support our research questions. Therefore, we used Grounded Theory [8] in our exploration to develop a model that captures such a behavior.

In the confirmatory study (Section 3), we refined and partially tested the developed model, by conducting an online survey among 397 Mechanical Turk (M-Turk) workers. The goal was to identify prominent factors that highly impacted users' decisions in practice.

Based on our findings, we offer guidelines on designing FOSN interfaces for reviewing and responding to friend requests (Section 4). While defending against large-scale infiltration is challenging [7], we hope that progress in this research direction will lead to the improvement of existing security defences and make them less vulnerable to both human exploits (i.e., automated social engineering [15]) and technical exploits (i.e., platform hacks [26]).

To summarize, this paper has the following contributions:

1. We developed a model for online lifecycle of Facebook friendship acceptance, which explains the factors that influence

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

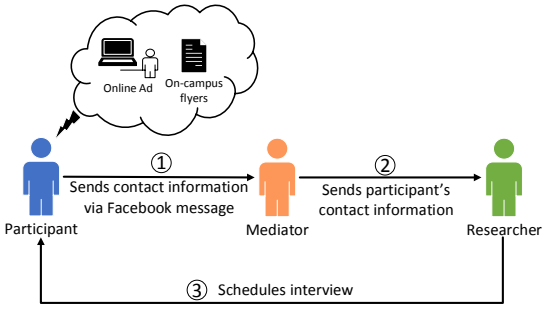


Figure 1: Mediator role

users' behavior in response to friend requests.

2. We characterized such factors and analyzed their impact on users' decision with regards to friend requests. We also identified four factors that significantly impact users' befriending decisions.

3. Based on both qualitative and quantitative results, we suggest design guidelines for FOSN interfaces that we expect can help users make informed decisions about friend requests.

2. EXPLORATORY STUDY

The study was in the form of semi-structured interviews. In what follows, we give more details about the study, including research questions, recruitment procedure, data collection and analysis.

2.1 Grounded Theory

We chose Grounded Theory as the approach of this study as it is an appropriate method for research in areas that have not been previously explored, especially when a new perspective might be beneficial [24]. Among different ways to apply Grounded Theory [13, 10, 8], we chose to follow the definition proposed by Charmaz [8] because it provides a more flexible format for data analysis.

2.2 Research Questions

In the exploratory study, we aimed to understand users' befriending behavior in response to friend requests, and to explore the factors that impact their decision. By applying the procedures of Grounded Theory coding, we were able to find new information, concepts, themes, and categories to develop a theoretical model, which helped in answering the following research questions:

- **RQ1:** What are the factors that influence users' decisions when responding to friend requests in general, and to friend requests sent by strangers in particular?

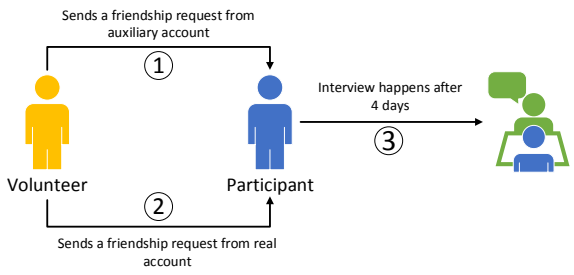


Figure 2: Volunteer role

- **RQ2:** What are the actions the users take before making a decision about a friend request?

- **RQ3:** What are the actions the users take after making a decision about a friend request?

2.3 Participant Recruitment

We posted the recruitment notices on local Craiglist and Kijiji websites. We also distributed flyers across our university's campus. In the recruitment notice, we included a brief description of the study and a hyper-link to an existing Facebook profile, and asked potential participants to send a personal message to that profile describing their interest, along with their email addresses.

We asked potential participants for their email addresses so that we have a reliable way to communicate urgent messages without depending on Facebook (e.g., unplanned changes in the interview schedule).

The owner of the profile was a graduate student in our department who was not affiliated with our research lab and was recruited to mediate the initial communication with potential participants. The purpose of recruiting a third party (i.e., the mediator) was to avoid any potential linkage between the user profile used for recruitment and our study. The mediator signed a non-disclosure agreement stating that all data collected through mediation would be immediately erased after relaying them to us, and that all information about the study would not be shared externally.

Overall, the mediator, denoted by M, operated under the following protocol, as illustrated in Figure 1:

1. A potential participant P uses Facebook to send a personal message to the mediator M, which contains P's email address and interest in the study.

2. M sends to the dedicated researcher an email including P's Facebook user identifier along with P's email address.

3. Once the researcher receives the email from M, he asks M to permanently delete the message that was sent by P and not to respond to any interactions initiated by P.

Using the email addresses of potential participants, we used e-mail to schedule interviews with them. We used the mediator to avoid inaccuracies due to self-reporting, when it came to identifying which of our participants tend to accept friend requests from strangers. This is why we had another volunteer who sent prospective participants friend requests from two other dedicated Facebook user profiles. The first user profile was a real account managed by another volunteer, while the second one was an auxiliary account that we created for the purpose of this study.[1] We aimed at reducing the chances that the participants knew the real account. To this end, we excluded students in our department from participating in the study.

As illustrated in Figure 2, the volunteer controlled both accounts and sent friend requests to potential participants according to our instructions. The volunteer, who was a graduate student from our department but not affiliated with our research lab, signed a non-disclosure agreement that prohibited him from both interacting with potential participants and sharing any collected information.

To avoid any suspicion among the participants in regards to the volunteer's account, we asked the volunteer to remove Facebook friends made for the purpose of the study after the interviews were finished, rather than before the interviews. While there was a risk

[1]The auxiliary account represented a male graduate student attending our university. The profile included a publicly available, generic picture of a man in his mid 20's.

of two participants having a pre-existing social connection (either online or offline) and seeing that the one is a friend with the volunteers, which could have influenced the other participant, none of the interviewed participants indicated that this was the case.

After each interview, we sent a debriefing message via Facebook to thank the participants for their interest in our study and provided them with more details about our research.

2.4 Data Collection

Our interviews were semi-structured, which gave us the flexibility to adjust and add new questions. We performed data analysis concurrently with the interviews in order to inform each new interview with the results obtained from the previous ones.

Each interview followed roughly the interview guide reproduced in Appendix A and had the following 6 parts:

1. Overview of the project.
2. Participants' demographics (e.g., age, gender, education, occupation, language) and Facebook usage-related questions (e.g., membership time, frequency of usage).
3. Participants' befriending behavior in general, and their responses to friend requests in particular. For instance, we asked questions about participant's friends, factors or criteria they employ to make a decisions about friend requests.
4. Participants' attitude towards their privacy and security.
5. Participants' attitude towards befriending strangers, and whether they had befriended strangers before.
6. Debriefing participants and concluding the interview. During this part of the interview, we also informed them about the friend requests that our volunteer sent. We observed each participant's reaction and asked each participant who accepted any of the two requests why they did so. We also asked participants if they had any suggestions regarding the interface design that might help them make more informed decisions.

As an iterative process, we analyzed the data by searching for patterns and forming concepts that were gathered into categories. We also wrote memos during the process of analysis to capture our understanding about the emerging categories and relationships among them.

Thanks to the iterative data analysis performed between interviews, we were able to detect "theoretical saturation" [14]. After 15 interviews, as Figure 3 shows, we reached the plateau where further data collection did not add new categories. This is why we stopped data collection after interviewing 20 participants. Their demographics are summarized in Table 1. All interviews were conducted in person at our university's campus. Each interview took about 50 minutes on average.

2.5 Data Analysis

As specified earlier, we employed Grounded Theory for the exploratory study. In Grounded Theory, data analysis involves searching for the concepts behind the answers. We transcribed, anonymized, and analyzed the collected data after each interview with an average turn-around time of 4 days. We used a web application tool called Dedoose for the analysis [1]. In what follows, we describe each part of the analysis in detail.

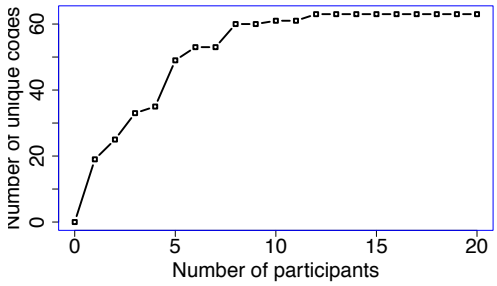


Figure 3: Theoretical saturation of interview data

2.5.1 Open coding

As the first step of coding, we identified, named, described, and categorized phenomena found in the collected data. Open coding resulted in a set of 63 unique codes, including both abstract (e.g., befriending behavior) and concrete labels (e.g., Facebook frequency of use). The intuition behind having abstract labels was to help develop a model. At the end, we had in total 2,620 coded excerpts, with an average of 131 per interview. We performed triangulation by having two other coders on four of the interview transcripts (interviews numbers 2, 6, 8, 11). The codes generated by the other two coders turned out to be subsets of codes generated by the main coder.

2.5.2 Axial coding

After open coding, we started to relate the generated codes to each other and ended up with 7 categories grounded in the collected data. The categories are *friendship factors*, *privacy and security awareness or concerns*, *investigation actions*, *decision execution*, *maintenance actions*, *environmental factors*, and *interface capabilities*.

2.5.3 Selective coding

The aim of selective coding was twofold: (1) to identify the main category, which ended up being **decision making process for friend requests**; and (2) discarded all categories that were not related to the core category, e.g., *fancy interface features*. Finally, we read the transcripts again and selectively coded any data related to the core category.

Demographics Type	Range	# of Participants
Age	19-29	11
	30-39	6
	40-49	2
	50-59	0
	60-69	1
Gender	Female	12
	Male	8
Facebook Membership (years)	0-2	7
	2-4	9
	4-6	3
	6-8	1
Facebook Friends	0-100	6
	100-500	9
	500-1000	5

Table 1: Demographics of interview participants

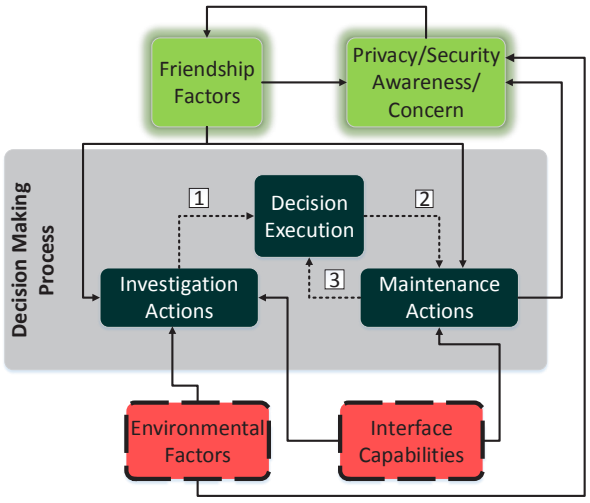


Figure 4: Online Lifecycle of Facebook Friend Acceptance (OLFFA) model. Shaded components on the top are the internal factors and components with hyphenated borders are the external factors. The middle box, which includes 3 components, represents the decision making process. The dashed arrows represent decision making flow. The solid arrows represent the impact of components on each other.

2.5.4 *Theoretical coding*

During this stage of analysis, we applied to the data the developed theoretical model. We integrated the model into related data in order to explain the core category. The outcome was a grounded model, or theory, about the lifecycle of Facebook friend acceptance, which we discuss in the following section.

2.6 Results

We now present the results of our exploratory study. First, we start by discussing the overall model, and then continue with detailed descriptions of the model components and the relationships among them.

2.6.1 *The Overall Model*

We refer to the developed model as the **Online Lifecycle of Facebook Friend Acceptance** (OLFFA). It includes 7 components, as shown in Figure 4. Each component is derived through the coding steps that were described earlier and is representative of a set of users' behaviors.

The factors that we found to have influence on the process of users' decision making can be categorized into four groups, to which we refer as components: Friendship Factors, Privacy and Security Awareness and Concerns, Environmental Factors, and Interface Capabilities. Since the first two components (green shaded rectangles in Figure 4) are user-specific and subjective, we considered them as *internal* (to the user). On the other hand, since a user does not have any direct control over the last two components (red rectangles with hyphenated borders), we call them *external* factors. The components inside the large grey box in the middle of the figure represent the decision making process, and the numeric labels indicate the flow of actions associated with decisions. The rest of this section discusses each of the components and the relationships among them.

2.6.2 *Friendship Factors*

This is the component that was brought up and discussed by all of the participants. Friendship Factors impacts Privacy and Security Awareness and Concerns of users in the sense that when users employ more restricted friendship factors, they become more sensitive about their profiles' privacy and security.

On the other hand, Friendship Factors could be impacted by Privacy and Security Awareness and Concerns. This happens when the Friendship Factors that the users employ change due to a an adjustment of their view on their profiles' privacy and security:

> "Well, from the time my brother's account on LinkedIn was hacked, I have always concern to have my info available on the internet. So I started to accept people that I feel comfortable to share my info with them. Not like before that I was accepting almost everyone." (P9)

As the result, a user could become more conservative in making new friendships. A reverse change could happen as well.

This component also impacts Investigation Actions and Maintenance Actions. For instance, if a user relies on the similarity of backgrounds for making friendships on Facebook, an investigative action could be to check out the requester's profile in order to see her background. Similarly, finding and removing passive friends is another example of maintenance actions driven by friendship factors.

Here is the list of Friendship Factors we have discovered:

- **Knowing the person in the real world (KRL):** It was reported by participants that they care about knowing people in real world or at least in online communities (e.g., forums), when they consider accepting friend requests on Facebook. For instance, P5 said:

  > "If I do not know them, I do not accept them. I mean I should have seen a person at least once to accept them as Facebook friend."

- **Profile picture (PRP):** The profile picture is one of the most important factors for users. We encountered users who usually spend only a few seconds to decide about friendship requests. Those users pay attention to only the profile picture, as the fastest way to make their decision. As P4 puts it:

  > "I can really know from pictures. If you do not have a picture then I do not know you!"

- **Profile name (PRN):** Similar to profile pictures, the profile name is used by users especially for the case when they want to instantly decide about friendship requests. They prefer to receive requests from recognizable names, to facilitate the process of decision making.

- **Common background (CBG):** During the interviews, many participants mentioned common backgrounds and interests as friendship factors. Users tend to accept friend request from people who have common background with them. These commonalities include city and country of birth or residence, schools and universities attended, personal interests, and hobbies, etc. When we asked for the reason, the users pointed out that these commonalities work like a trigger that helps them remember the people they have on Facebook and to know them better. For example, P17 said:

  > "Although it is fine for me to have new friends based on my interests, I would prefer to be in the same city to make closer friendships."

- **Being active on Facebook (BAF):** According to our data, the fact that the friend requester is an active Facebook user is sometimes the most important factor, even more than knowing the requester. P5 expressed this by saying:

> "If they send me a request, okay, I know you. I am going to accept your request but it has been five months and you are not posting anything. You never come to Facebook. You never post anything. Okay, I am sorry. I have to delete you because you are not adding anything."

- **Gender (GEN):** The gender was another factor for participants. P5 said:

> "I think gender is effective in terms of friend requests. You know, I am sorry to say it but put a picture of a pretty girl would get hundreds of friendship requests or even messages. I have a male friend who was building a 'stable' of Facebook women. He had about 600 friends and they were all women. There is not a single male friend on the list!"

- **Number of mutual friends (NMF):** The majority of participants confirmed that the number of mutual friends is important, as it helps users to remember whether they know each other. Although it is known as a way of verification by many users, it might fail them. P2 raised an interesting point about it:

> "I used number of mutual friends as a fast approach to accept friends but later it turned out it is not necessarily good enough because I removed many friends who had large number of mutual connections with me. Maybe because I had a lot of friends, around 800, so I had many friends in common with people and it did not work all the time."

- **Closeness of mutual friends (CMF):** Some participants highlighted that, in addition to the number of mutual friends, it is also important to know the closeness of those friends. That is, even if there are a couple of mutual friends between the receiver and the requester, it is not necessarily enough for users to make a decision. As P5 expressed it:

> "You either have to be someone I know or you have to be mutual friends with someone I really know. Anyone else I do not take requests anymore because I ran into some pretty weird people."

- **User's activity pattern (UAP):** Another friendship factor was user's activity pattern, including what kind of information is shared (i.e., either relevant or irrelevant) and how often the content is shared. For instance, P1 said:

> "I do care about what they post. If they post, like, things that I would find disturbing for me, ding!! I would delete them."

Furthermore, our participants disliked being friends with those who just monitor others' posts, and possibly report to mutual contacts:

> "My aunt turned out was watching my page and then reported my activities to my mom. And that did not go over well and I just blocked them. I would never befriend anybody who just monitors others." (P6)

Given this dislike for passive users, it was interesting to discover that some of our participants had changed their activity on Facebook over the years. They undergone a shift from active to passive users, who just read others' posts, without regularly adding any content. According to our participants, an active user is the one who is willing to have a lot of Facebook friends and performs a variety of activities, such as sharing photos, notes, and videos, as well as posting their status, etc.

- **Closeness and quality of friendship in real life (CFR):** We found in the interview data that it is important for users to make sure how good of a friend they might become with the requester and if they might get along. For instance, P6 reported:

> "If I know them then, it takes a little bit longer because then I have to decide because my half-brothers and their daughters have requested to be my friends. And yes, I know them but, no I do not want them on my page. Because the girls I do not get along with when they come for Christmas dinner. We only see them at Christmas time and I do not get along with those girls. My half-brothers, the one I do not – I have only met this past summer for the first time, so I do not know him and I am not interested!"

Another participant, P5, expressed similar concerns:

> "I found this quite upsetting but there is a woman on my site who I worked with. We were quite close at work but I did not like a number of things that she did, and you know I did not accept her request."

- **Application-based friendship (APF):** There was another factor raised by our participants where users tend to make friendships with others for the sake of receiving bonuses from some applications such as games. As a result, such users would send and accept more friendship requests.

2.6.3 Privacy and Security Concerns and Awareness

As described earlier, this component is influenced by and impacts Friendship Factors. Maintenance Actions also impacts this component. This might happen as a maintenance activity, for example, when a user monitors a friend's profile and she ends up facing surprisingly irrelevant content posted by this friend. This observation would cause them to be aware of fake or hijacked accounts posing as close friends:

> "I remember that I found that there were two accounts for a friend of mine and I thought he had created another one. When I asked, it turned out that the first one was a fake account and he had already deactivated his previous account. So, somebody had created an account similar to his first account. I did not know that. I even checked my name to see if there is any fake account for me as well as other friends." (P17).

Another source of influence on this component is Environmental Factors in general and media in particular. Some participants noted that their awareness of privacy and security on Facebook were affected by media reports. For example, P7 shared:

> "Previously, I would just add like a lot of random people and accept requests. Later, I became more conservative, as I heard from media about leakage of users' information."

P1 also believed that there were security incidents reported by media that influenced her behavior:

> "Because there are a lot of issues with Facebook, like pictures, as there was the recent one about the girl who committed suicide and how her photo was used for some porn website so things like that. So for the pictures that I post on Facebook, they are never of my face."

P3 had similar concern describing his experience:

> "I used to post a lot of photos on Facebook but then there are issues with security. The more you post, the more you cannot take back because I read in a blog that even if you post a photo on Facebook and get rid of it from your account, just delete an album, you are still going to be on Facebook. So because of that I stopped posting photos on my account."

We also found an interesting point about the effect of security and privacy incidents in other online services, which results in change of behavior on Facebook. P10 said:

> "I had profiles on LinkedIn and Evernote but then I removed it because of some security leak in passwords. I got sensitive in terms of disclosing information on my accounts."

2.6.4 *Interface Capabilities*

Our participants reported a set of issues related to capabilities of the interface—e.g., lack of required information, device-specific design, and frequent changes of privacy settings—that would impact Investigation Actions and Maintenance Actions.

Some of the participants could not easily find desired information in order to make decisions about friendship requests. As a result, they preferred sometimes to think about requests, rather than looking for additional information on Facebook about the requesters. This raises the issue of information visibility in the interface. For instance, P3 provided the following suggestions:

> "Definitely need to have what/where they are from, what they have, if it is in academic backgrounds, then what they studied and where. And if it is just maybe a few interests that they have, [it] could never hurt, I think. Just because you look at a person and you think they are interested in photography I do not think it could actually hurt anyone. So just something along those lines that can give you more information."

Regarding the issues related to device-specific design, P8 shared her experience as follows:

> "In terms of an interface, maybe a bigger button, I think just because sometimes all those buttons look very similar and you tend to click one. If you are using your phone and looking at someone who you are not a friend of, but you want to (this has happened to me before), you want to message that person instead before you add as a friend and then by mistake because the buttons are right next to each other I would press add a friend, send a friend request, or add a friend instead of message. So when that goes out that is it. They receive it and then you cannot really retract that."

P13 mentioned another issue in this regard:

> "It really depends if I use my phone or my desktop when I accept or reject a request. Using the desktop, I spend way more time while this is not the case with my iPhone. So you would be lucky to have me on desktop when receiving your request. On iPhone, I would make my decision very quickly. If I do not remember, I would just reject."

This issue shows the gap between usability of device-specific designs of interfaces for accepting/rejecting requests.

The last issue about the interface was frequent changes made to the interface, the privacy settings in particular. Participants found it difficult to catch up with these changes.

2.6.5 *Investigation Actions*

Before making their mind in regards to friendship requests, some of our participants took one or more of the following actions:

- **Sending personal message:** Specified by many participants, sending personal message is a common technique for obtaining additional information about the requesting user, especially when he is not known to the receiver. As P7 explains:

  > "I would personally ask them on private messaging and say that I do not know you or asking some questions like 'have I met you?' "

- **Checking out photos:** It was also common among the participants to go to the profile and, if possible, check out photos of the requester. They reported to be helpful to recognize the requester, to either make decide about the request or start communicating with the requester via messaging.

- **Looking for commonalities:** Another action taken by our participants was to explore for commonalities in terms of background, friends, interests, etc., as P5 illustrated:

  > "Do we have common interests? Do you know some friends of mine? We have something in common maybe?"

  This action seemed to be done by those participants who had new friends, in order to help them know people better, as well as those who wanted to have limited list of friends, in order to help them verify requesters, in case the profile picture or name were not recognized.

- **Checking mutual friends profiles:** Some of our participants reported that, although it was important to know if there were any mutual friends, it also took time to check out the mutual friends' profiles for evaluating the closeness of the relationship. Although it was important to some of our participants, some other participants said that they would skip this step because it was too time-consuming and required somewhat high cognitive load:

> "I really want to know more than just number of our mutual friends and see if those are close friends but I check that when it does not take me a long time. Like less than 5 minutes otherwise I won't do that." (P13).

2.6.6 Decision Execution

We found three types of behavior for decision execution. (1) Some participants would make their decisions immediately after they received requests. If they could find information they needed to make the decision, then they would easily make it right away. There were other participants who would accept friend requests right away, although for different purpose. They would do so in order to find out more about the requester (after becoming friends) and then decide if they wanted to unfriend her or not.

(2) Otherwise, they would reduce their set of decision criteria, in order to expedite the process. In such cases, participants with less concerns about privacy and security would most likely accept friend requests:

> "If I get a friend request that we share mutual friends but I do not know them, I am always hoping that I can check their profile. Sometimes it is restricted so you cannot. So I accept the friend request." (P5)

(3) On the other hand, some users would leave requests as they are, and postpone further investigations.

2.6.7 Maintenance Actions

The interview data revealed three types of Maintenance Actions that our participants took after accepting friend requests.

One of the common maintenance actions was to remove friends after a while, due to a number of different reasons. For examples, those friends that had been added in order to play face boo games, would be removed when there was no need to be friends with them. Another common reason was finding content shared by to-be-removed users irrelevant. As a result of these actions, users may adjust their Privacy and Security Awareness and Concerns, which would eventually impact their Friendship Factors.

One other type of maintenance actions was to define different levels of access for friends. This usually happened in two ways. One was to define separate groups of friends and then specify visibility of the posts using these groups. The other way was to deny specific users the ability to see a post or any desired content on-the-fly. This means that participants sometimes set the access level manually to avoid a group of friends accessing the post. As an example, P7 said:

> "If it is for family pictures, I would just change the privacy setting to relatives. Then, I do not have to remember every one of those friends. Sometimes I do not even have to create a group for relatives though. I can remember who are my relatives."

The third type of actions was for our participants to update the privacy settings of their profiles. However, some of our participants, who were sensitive about their privacy, complained about frequent changes that Facebook privacy settings undergo:

> "It changes a lot, but from time to time I try to go back and look at it, but that could be like once a year or so." (P3)

On the other hand, we found that some participants were not even aware of privacy settings in the interface. When we asked about the possibility of access to information of their profiles, some of them did not even know if it were possible. P2 said:

> "I guess so, because I have not seen that at all. But, now that you have talked about that, to me that means there are thousands of people that can check who I am. Some groups are pretty big. I have not thought of it."

This issue with frequent changes in Facebook privacy settings illustrates the relationship between Maintenance Actions and Interface Capabilities, in which the latter impacts the former.

2.6.8 Environmental Factors

Analysis of interview data revealed that there are three environmental factors that influence Investigation Actions and Privacy and Security Awareness and Concerns, as discussed before.

First, the participants referred to the lack of time, as a factor that influenced their decisions about friend requests. For instance, P17 said

> "I have always problem with the lack of time during break times. I have to check updates, requests, messages, etc. in just 15 minutes. I once accepted a friend by mistake, as the requester had just same name as a friend of mine and I had not checked his profile to get more info about him."

The second factor is the lack of concentration, while checking out Facebook:

> "On the way to university, I usually check out my profile on the bus. I once accepted a request when I was on the bus and that was a wrong decision. I guess I was distracted by stops and also other passengers so that I forgot to send a message to the requester." (P20)

The third environmental factor was the effect of media. As described earlier, the Privacy and Security Awareness and Concerns of our participants were impacted by media reports about security and privacy incidents.

2.7 Discussion

In order to answer the research questions, we decided to go one step back and envision the problem as part of a bigger context. Therefore, we managed to come up with a model which discusses users' behavior when they want to accept/reject a friend request. This idea was supported with the fact that there is no previous study focused on this aspect of users behavior. Armed with such a model, we would be able to uncover behavior of users towards strangers since this scenario would be a specific case of the model. We define stranger as a person who is not familiar in real life or online communities. In this regard, we indirectly asked participants about their interaction with strangers so that we can reveal more details about this scenario.

2.7.1 Befriending Strangers

As described in Section 2.3, before each participant was interviewed, the participant received two friend requests, one from a Facebook profile of a real user, and the other from an auxiliary profile made up for the purpose of the study. Five participants accepted at least one request from one of these accounts, and one of them accepted requests from both accounts. When we reached in our interviews the debriefing part, in which we informed the participants that these requests were from our research team, their reactions varied.

The participant who had accepted both requests said that it was okay with him and he did not care about strangers among his Facebook friends, since he did not have any idea that anybody could

make any use of his profile data. The other four participants who had accepted requests from either real or auxiliary accounts of the researchers had different attitudes. After hearing the scenario, they got nervous and one of them said:

> "I would not have accepted the request if I knew more. I saw the guy is from UBC and is a graduate student. I thought that it should not hurt."

Another participant, most of whose profile was accessible publicly, had similarly nervous reaction, especially when we explained the possibility of any user accessing his profile information. He commented that in the future, he would pay more attention regarding friend requests.

In addition, we found evidence in interview data suggesting that some OSN users don't pay attention to possible threats, when it comes to making friendship connections:

> "I seem to be a million times more strict than most people. I know some friends who accept anybody that requests. Well, I mean a lot of people do. They do take it too easy. How can you have 2,000 friends?" (P5)

Another participant had a set of "friends" from accessory shops (she did not know them) while they had access to the profile information e.g., other friends in her profile. Some participants seemed to have no criterion for making friendship. They would just add anybody, as P11 explained:

> "I am always nice to requests on Facebook, as I cannot remember that I have rejected a request."

Attitudes Towards Strangers: These observations made us more curious about users' perception of Facebook users they do not know in real life. Our analysis suggests that, when it comes to one's attitude towards strangers on Facebook, our participants can be roughly divided into three groups.

We found that one group of participants had a "take it easy" attitude towards accepting friend requests from strangers. As P1 justified:

> "I have spent some time with them on Facebook and they do not seem somebody who would cause me pain!"

As P1 mentioned, it is enough to have a feeling that a person is not going to make any trouble for them. The other reason for accepting their requests is that having less commonality might be even an advantage, as P16 illustrated:

> "I know some people in real life who have common things with me like our neighbor's kids that we lived in the same neighborhood, we went to the same school. But I do not want him to be on my Facebook profile. I prefer to have more of these unknown guys instead of our neighbor's son, as some of them post cool stuff and I don't need to be worried about my posts, because none of them would tell my dad what I am doing!"

On the other hand, for some other participants, only knowing a requester in real life did not necessarily mean that this was a right person to be friends with on Facebook. P2 illustrated this point with the following example:

> "I have like friends from primary school who ask me to be [Facebook] friends. But, in primary school you are friends with all your classroom so then it will be like your real friends. And that has not been done for 15 years. So now I do not accept them anymore if I see that we are in really different world and everything. It is my private life and I am a new person now."

P1 explains this attitude further:

> "If you have not kept in contact or you have not actually tried to stay in contact, I feel like there is no point. Long ago in the past, I do not go back there."

Users who have this attitude are less vulnerable to the threat of accepting a stranger's request.

The third group's attitude was not as clear cut as for the first two groups. As a result, participants from this group were influenced by the various factors specified in our model. This group would be also vulnerable to the threat of accepting strangers' requests, as participants from this group reported issues in recognizing people in real life or online communities.

These groups are not necessarily mutually exclusive, i.e., the same user can exhibit in the majority of cases the behaviour of one group, and yet handle some of the requests following the pattern of another group.

Accepting While Not Indending: Our analysis revealed that some of our participants would make inconsistent decisions. For instance, they would accept friend requests although they didn't have intention to be Facebook friends with the requesters, as an example of P11 illustrates:

> "Some requests are from people that I had a quick chat with them or somehow I remember them but honestly I don't want to be friends with them. However, I will accept if they send me request."

These participants seem to find it socially awkward to reject friend requests. P18 made it explicit.

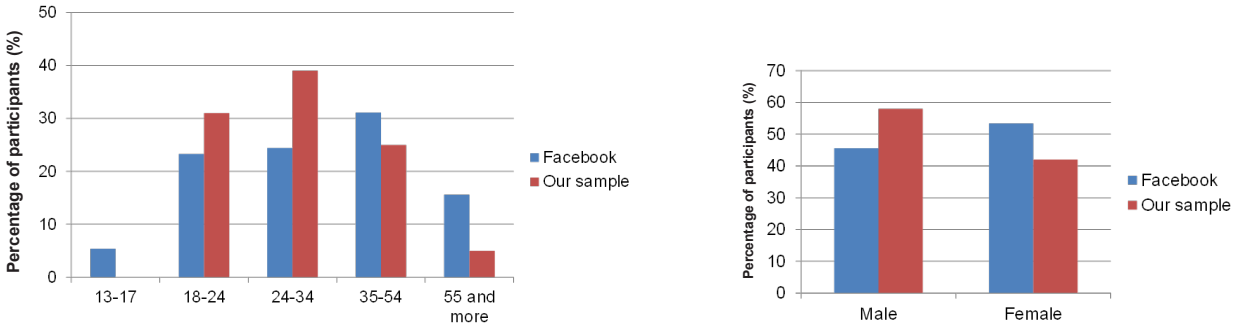
> "I always have this problem with some of people I know but I don't have a really good relationship with them that I cannot say no to their request. I don't know why but I think it's better to accept rather than reject them." (P18)

Usage Differences: We discovered differences in the way our participants used Facebook, and these differences seem to correlate with they way they treated friend requests. Although it has been previously shown that users tend to use OSNs (including Facebook) to make connections and share different kinds of data, we found three "flavours" of users:

- **Contributors**: These are traditional users who both consume and contribute new content. They make friendships, share photos, share personal information, post updates, and interact with others by commenting and favoring their shared content. From the point of view of this group, the aim of FOSNs is to make an environment in which people feel free to share information with others and receive feedback. While they are willing to have more friends, they are also conscious about their profile privacy and friendship management, as P16 illustrated.

  > "I really enjoy using Facebook when I share posts or comment on a post and receive likes. But this is because I know my friends and feel comfortable with them"

- **Observers**: On the other end of the spectrum, there are users that avoid having social interaction and prefer to passively observe others. They have different reasons for this behavior including lack of time, security concerns, difficulty to use the interface. As the result, they do not share any information and they are willing to make connection with as many users as possible.



(a) Age

(b) Gender

Figure 5: Age and gender comparison of our sample to Facebook population.

> "I like Facebook as it gives me the chance to read my friends' posts and watch their photos, read news and many other things. Of course I don't share anything as I use my phone and it's really difficult to type a lot. Moreover it takes a lot of time." (P13)

- **Conscious Contributors**: In addition to these two extremes of the spectrum, there are advanced contributors who are more sensitive about the audience of their posts and other shared content. This third group of people reports more issues regarding friendship management, as P15 illustrates:

> "What I am looking for on Facebook is to interact with others and share my info as well as see their posts. I am spending a lot of time to manage my profile and I have this difficulty to put my friends in different groups as I want to have them but I don't like to share my personal photos or posts with all of them."

To summarize, our observation indicates that we can categorize users of FOSNs into three groups, with Contributors and Conscious Contributors being more likely to have issues in terms of privacy and security of their profiles. This sheds light on the point that privacy and security would have different meanings for users according to the type of their FOSN usage. Consequently, this may impact user's attitude towards friend requests.

Our Online Lifecycle of Accepting Friends model could be helpful for FOSN designers, when it comes to supporting users in deciding about friend requests. The model could aid in considering various factors that impact user decisions.

3. CONFIRMATORY STUDY

While the exploratory study allowed us to identify possible factors that have a role in users' decisions about friendship requests, we wanted to test these factors on a representative sample and measure the fraction of users who are employed by those factors. Therefore, we decided to conduct an online survey that would allow us to collect quantitative data from a representative sample.

For each of the eleven friendship factors identified from the interviews, the survey had at least one statement (e.g., "If I recognize someone's picture, I would accept his/her friendship request on Facebook.") and asked participants to indicate their agreement on Likert scale of 1-5. For those factors that had more than one statement, we used the mean score. For testing data quality, we have included contradicting statements. For example, "I would accept a friendship request from a Facebook application." and "I don't tend to accept friendship requests sent by Facebook applications." All questions from the survey can be found in Appendix B.

We recruited 425 M-Turk participants from USA and Canada. Each USA participant received $0.50 and Canadian $0.75. It took 16 minutes on average for our participants to finish the survey. We removed 28 participants because of contradictions in their answers, which left us with responses from 397 participants.

3.1 Results

First, we provide statistics related to sample representativeness and participants demographics, then descriptive statistics regarding employment of the friendship factors, finally we discuss the impact of the friendship factors on accepting a stranger's request.

3.1.1 Pariticipants Demographics

We compare demographics of our sample with the demographics of Facebook users.

As Figure 5a shows, our sample is younger than Facebook users. We got more younger participants (18-24: 31% vs 23.2% and 24-34: 39% vs 24.4%) and fewer participants in higher age ranges (35-54: 25% vs 31.1% and 55 and above: 5% vs 15.6 %). We did not have any preference to recruit participants from younger age range and as mentioned earlier, we recruited participants from Amazon M-Turk. However, previous work shows that the turkers are relatively young with about 80% in 18 to 40 years old age range (Average = 31, Minimum = 18, Maximum = 71, Median = 27) [22], which could be the reason for having a younger sample rather than Facebook demographics. It is also worth mentioning that we did not have any participants in the age range of 13 to 18, as we chose to recruit participants who were at least 19 years old.

In terms of gender, as Figure 5b shows, our sample was biased towards male participants (58% vs 42%), while 53.3% of Facebook users are female and 45.7% are male.

Demographics of our participants show diversity of the sample. In terms of age, we had participants from 19 years old to 65 and more. Gender-wise our participants were fairly evenly distributed. Participants also had diverse education levels (26% with high school or lower degree, 59% with undergraduate degree, 10% with graduate degree). The employment status of our participants varied, too: 56% employed, 22% students, 16% unemployed, 2% unemployed and 4% had other employment status.

We also asked our participants general questions about their Facebook usage and experience. The majority (94%) were Facebook

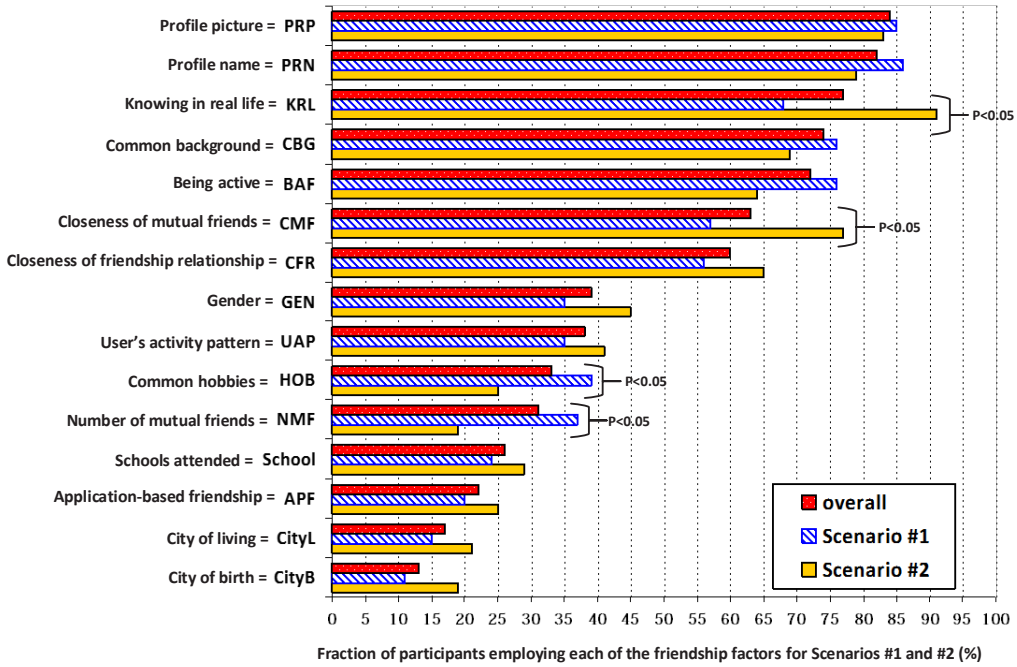


Figure 6: Distribution of friendship factors employment among all participants, scenario 1 (S1), and scenario 2 (S2). Significant differences between participants of S1 and S2 are shown in terms of employing KRL, CMF, NMF, HOB ($p < 0.05$).

users for more than 2 years. In terms of usage frequency, 92% reported that they login into Facebook at least once a month, while 80% login several times a week. They were also asked to go to their Facebook profile and enter the exact number of their friends. Our participants had wide range of friendship circles, with minimum of 10 and maximum 3,000 (mean 328, median 203). This shows that collected data came from users with different befriending patterns. Majority (64%) of participants receive at least one friend request in a month and only 7% receive friend requests less than once a year.

3.1.2 Friendship Factors

Figure 6 summarizes results of the survey on the friendship factors. The red bars show the percentage of all participants who reported employing each of the factors, i.e., they agreed or strongly agreed with the corresponding statement(s).

Starting from the most popular factors, requester's profile picture (84%) and name (82%), participants accept friendship requests if they recognize the requesters. Seventy seven percent agreed with statement "I tend to accept friendship requests from people I know in real life or online communities."

Another factor was "common background" (CBG). While 74% of participants agreed that it is important to know requester's background, the survey results show that the participants were not specifically interested in a single type of background information. And the importance varied among participants. For instance, only 15% would accept friend requests from users who were born in the same city as they were. Similarly, only 18% would accept friendship requests from users who live in the same city as they do. On the other hand, 27% would be interested in having Facebook friends from the same school/university. The most popular type was "common interests/hobbies," with 35% relying on this background information in their decisions about friend requests. This particular result was corroborated in the interviews, with participants reporting interest in new FOSN friendships with those who share interests or hobbies.

Another factor that we tested was activeness of friends, with 72% reporting interest in accepting friend requests from active users. In terms of gender (GEN), 39% of participants confirmed they consider it during decision making for friendship requests. The "number of mutual friends" (NMF), which is currently shown in the Facebook's friendship request dialog, was only used by 31% of participants for making their decisions. On the other hand, the majority of participants (63%) do care about "closeness of mutual friends" (CMF) to them. Regarding the impact of "user activity pattern" (UAP), we found that 38% of participants were reluctant to accept a friend request if they saw irrelevant posts shared by the requester. This was expected, as our interviews showed that although people like to have access to the posts of requester, they usually do not have this level of access. The results also show that "closeness and quality of friendship in real life" (CFR) was important for 60% of participants. We also measured the number of participants who would accept "requests from Facebook applications". Results show that 22% of participants took APF into consideration, as a factor in deciding about friend requests.

3.1.3 Accepting Friend Requests from Strangers

We wanted to understand if there is a difference between those participants who accept friend requests from strangers and those who don't. We were specifically looking at the difference in the way they would be influenced by the Friendship Factors.

To investigate this difference, we considered two types of user behaviour, which we describe as two scenarios: (1) **(S1)**: users could accept friend requests from strangers, and (2) **(S2)**: users would reject friend request from strangers. In these scenarios, friendship factors are dependent variables (DVs) and a decision of either accepting or rejecting friend requests is the independent variable (IV).

We divided our dataset into two groups (scenario 1 and 2). This was done by analyzing the answers to one of the survey questions, which explicitly asked participants if they have any strangers among their Facebook friends. 62% of the participants confirmed that they did. Then, we compared these two groups in how much they used each of the friendship factors. In what follows, we describe the results of our comparison.

We found that while only 68% of participants in S1 consider the knowledge of the requester in real life (KRL) in their decision process, this number jumps to 91% for S2, with th difference being statistically significant (Mann-Whitney's test: $p = 0.0003 < 0.05$). We interpret this result as an indicator for the level of awareness in these two groups.

For profile name (PRN), although we did not see much difference between the groups, participants in S1 reported more interest than those in S2 (80% vs 87%) for using profile name as a factor.

For common background, we looked at four types of background information, including city of birth (CityB), city of Living (CityL), schools/universities attended (School), and common hobbies/interests (HOB). For the first three factors, we could not find statistically significant difference between participants in S1 and S2. However, S2 participants were slightly more interested in them (CityB: 19% vs 12%, CityL: 21% vs 15%, School: 29% vs 25%). The difference was significant when it came to "common hobbies/interests" (HOB). While 40% of participants from S1 employed this as a friendship factor, there were only 25% in S2 who did so (Mann-Whitney's test: $p = 0.03 < 0.05$). This result could be leveraged as a cue by socialbots to customize profile information in order to increase the chance of getting their friend requests accepted. "Being active" (BAF) was also more popular among S1 (76%) members rather than S2 members (64%), although the difference was not statistically significant.

Regarding the "number of mutual friends" (NMF), we saw significantly more members in S1 (37%) than S2 (19%) employing it as a factor in their decisions (Mann-Whitney's test: $p = 0.01 < 0.05$). Also, comparison of S1 and S2 in terms of "closeness of mutual friends" (CMF) indicated that more participants in S2 (77%) cared about it than in S1 (57%) (Mann-Whitney's test: $p = 0.03 < 0.05$). The results of comparison for NMF and CMF suggest that informing users about the closeness of the requester with the mutual friends would be more effective than only showing the number of such friends (available in current interface).

For user's activity pattern, we found that participants from S2 were slightly more interested in UAP than from S1. We suspect that the absence of statistically significant results in regards to UAP is due to the difficulty of finding a pattern, as we had this feedback in exploratory study. Regarding closeness of friendship relationship, we did not find statistically significant difference between S1 and S2. This result is expected, as it more relates to scenarios in which friendship requests are sent from known users, according to our interview data. Finally, we could not find statistically significant difference between participants in S1 (20%) and S2 (25%) regarding application-based friendship (APF), although we expected to observe significantly more participants in S1 who rely on this factor. This might be because of the shortage in the number of participants who have received this type of friendship requests.

4. DISCUSSION

Considering the first goal defined for the survey, we analyzed the data related to each of the factors to investigate how much they are used. As the result, except for UAP and APF, all other friendship factors were employed by at least more than 50% of participants, which shows the validity of friendship factors inferred from the exploratory study. In addition, we asked survey participants to share with us other friendship factors if they have any. Analysis of answers to this question did not add to the factors themselves. The participants who answered this question, mostly suggested features that could be added to the friend request decision dialogues. As mentioned earlier, since having access to user's wall is usually not possible, people may not consider UAP as a factor. However, according to the exploratory study, participants prefer to have information about the activity patterns of requesters. For APF, a low percentage was expected from the interview study, in which only few participants reported receiving friendship requests from applications.

For the second goal, the idea of focusing on the results of groups who have strangers in their Facebook friends, and comparing it to those who do not have, helped us to investigate and uncover the impact of the friendship factors. As the results show, we found four friendship factors (KRL, HOB, NMF, CMF) could play a notable role and influence users' decisions. This result could be leveraged for improving the interface design so that users make more informed decisions.

4.1 Interface Design Recommendations

As discussed before, the results from the analysis of our survey data revealed interesting points about friendship factors that could be used for improving the Facebook interface. Therefore, we offer the following suggestions for designing user interfaces for accepting friendship requests:

- The interface should convey the importance of making accurate decisions about friendship requests and encourage users to make informed decisions. For instance, users could be notified by a pop-up window (similar to current design) asking users to go to another page in order to make an informed decision, using useful information or a check list. Having such a feature in the interface is supported by the OLFFA model since it helps users to appreciate the importance of these decisions.

- The interface could contain a message box so that requesters can briefly specify how they know the user. Another suggestion is to give access to photos selected by each user to better recognize the requester. We had reports from participants of both studies complaining about unclear small photos. This kind of improvement would facilitate the investigation/maintenance actions (in the decision making process of OLFFA model) for users.

- It could be helpful if user had access to statistics (number of likes, number of comments, number of personal messages, number of common photos) about interaction with his/her friends. In this case, it is easier to investigate closeness of mutual friends, which was shown to be more useful than only the number of mutual friends. In other words, this feature would facilitate the Investigation Actions in the OLFFA model for finding out closeness of mutual friends.

- The interface could encourage the user to specify the access level for new friends at the time the user accepts a friend request. We suggest this because our analysis showed that 31% of participants in S1 did not define any access level for their friends while 9% in S2 reported similar behavior. Therefore, this could be helpful (at least for users who accept stranger's requests) as a facilitator for performing maintenance actions and help users to be more cautious about the level of access they grant to their Facebook friends.

It is worth mentioning that although we believe these recommendations could be helpful for the Facebook interface improvement, they are currently hypotheses to be tested.

5. LIMITATIONS

Our work has several limitations. In the exploratory part, it would be better to have more diversity in terms of age so that the model could be representative of a wider range of Facebook users. On the other hand, although we reach saturation in data collection, we had five participants who accepted friendship requests from the volunteer. Having more participants from this group could result in more interesting observations and a more accurate model.

In the survey, we asked participants to report their activities, which might not be accurate due to somewhat abstract nature of the questions. As an alternative, it could be done by providing them with different scenarios and then asking them questions. We refrained from doing this due to the time limits of our survey. Finally, our sample is not representative of all Facebook users, as we recruited participants only from USA and Canada. Having participants from other countries could reveal more interesting points about users befriending behavior.

6. RELATED WORK

Previous work shows that changes in friendship network has been observed due to internet use. For instance, friendships continue to be abundant among a wide range of adult Americans from (25 to 74 years old) from 2002 to 2007 [27]. Emergence of online social networks was one of the main reasons for this phenomenon. While the number of OSN users is still growing, there are concerns about privacy of users. There is work on definition of privacy, and digital privacy in particular, to clarify what should be expected by users in terms of privacy [21]. On the other hand, it has been shown that this is not always a fault of systems that results in privacy and security issues and humans are a major cause of these failures [25]. Therefore, it is necessary to consider humans in designing systems. Cranor proposed a framework to reason about the human in the process of designing secure systems [11]. This framework was insightful during the process of qualitative data analysis to form our model. The is also work related to privacy of users on Facebook. It was shown that users' intention does not match with their privacy settings [18, 19]. Another study showed that users have difficulty in understanding the privacy settings and cannot configure them correctly [12]. As the most related work to ours, Johnson et al. showed that the main concern is insider's threat rather than the outsider's [17]. We believe that the focus of our work is different, as our concern is to understand user's behavior towards friendship requests rather than how they manage their privacy settings. Moreover, we believe that stranger's threat still exists as 62% of our sample reported to have at least one stranger in their friend list.

7. CONCLUSIONS AND FUTURE WORK

Our work contributes to providing socio-technical solutions to help users be aware of their decisions towards friendship requests from strangers. First, we aimed to better understand their behavior. We identified three groups of factors that impact users' decisions, including internal factors (Friendship Factors, Privacy/Security Awareness and Concern), external factors (Environmental Factors, Interface Capabilities) as well as a 3-step process of decision making (investigation, decision execution, maintenance). We believe that this model is helpful for improving the part of interface related to receiving friendship requests. We also showed that accepting stranger's requests is still a threat, as having at least one stranger in friend list was reported by 62% of our participants. We also introduced 4 friendship factors (knowing in the real world, common hobbies/interests, number of mutual friends, closeness of mutual friends) that can significantly impact users' decisions in regards to friend requests. Then, we offered suggestions for improving the interface.

There are several directions for future work. One direction is to perform structural model testing on the proposed model Structural Equation Modeling (SEM). Another direction is to conduct a user study and investigate impact of modifying the interface using the proposed guidelines. Another one is to focus on each component of the model and investigate their potential impact on friend request decisions.

8. ACKNOWLEDGMENTS

This work was supported by NSERC. We would like to thank LERSSE members for their constructive feedback on this work.

9. REFERENCES

[1] `http://www.dedoose.com/`.

[2] J. Bates. Sniffing out socialbots: The combustive potential of social media-based algorithms. `http://www.huffingtonpost.com/john-bates/financial-trading-algorithms_b_1125334.html`, December 2011.

[3] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on World wide web*, pages 551–560. ACM, 2009.

[4] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[5] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 93–102, New York, NY, USA, 2011. ACM.

[6] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. Design and analysis of a social botnet. *Computer Networks*, pages 1–22, 2012.

[7] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. Key challenges in defending against malicious socialbots. In *Proceedings of the 5th USENIX conference on Large-scale exploits and emergent threats*, LEET'12, Berkeley, CA, USA, 2012. USENIX Association.

[8] K. Charmaz. *Constructing Grounded Theory*. SAGE publications, 2006.

[9] E. Chung. Facebook easily infiltrated, mined for personal info. `http://www.cbc.ca/news/technology/story/2011/11/07/technology-facebook-socialbots.html`, November 2011.

[10] J. Corbin and A. Strauss. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage, Newbury Park, CA, 1990.

[11] L. F. Cranor. A framework for reasoning about the human in the loop. In *UPSEC'08: Proceedings of the 1st Conference on Usability, Psychology, and Security*, pages 1–15, Berkeley, CA, USA, 2008. USENIX Association.

[12] S. Egelman, A. Oates, and S. Krishnamurthi. Oops, i did it again: mitigating repeated access control errors on facebook. In *CHI*, pages 2295–2304. ACM, 2011.

[13] B. Glaser and A. L. Strauss. *The Discovery of Grounded Theory, Strategies for Qualitative Research*. Aldine Publishing Company, Chicago, Illinois, 1967.

[14] B. G. Glaser. *Theoretical sensitivity : advances in the methodology of grounded theory*. Sociology Press, Mill Valley, CA, 1978.

[15] M. Huber, S. Kowalski, M. Nohlberg, and S. Tjoa. Towards automating social engineering using social networking sites. *Computational Science and Engineering, IEEE International Conference on*, 3:117–124, 2009.

[16] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, 2007.

[17] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and privacy: it's complicated. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 9. ACM, 2012.

[18] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC '11, pages 61–70, New York, NY, USA, 2011. ACM.

[19] M. Madejski, M. Johnson, and S. Bellovin. A study of privacy settings errors in an online social network. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 340–345, March 2012.

[20] F. Nagle and L. Singh. Can friends be trusted? exploring privacy in online social networks. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, pages 312–315, Washington, DC, USA, 2009. IEEE Computer Society.

[21] L. Palen and P. Dourish. Unpacking "privacy" for a networked world. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 129–136, New York, NY, USA, 2003. ACM.

[22] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[23] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 249–252, New York, NY, USA, 2011. ACM.

[24] P. N. S. Rita Sara Schreiber. *Using Grounded Theory In Nursing*. Springer Publishing Company, ISBN 0826116221, 2001.

[25] B. Schneier. *Secrets & Lies: Digital Security in a Networked World*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 2000.

[26] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In *Proceedings of the 4th Workshop on Social Network Systems*, SNS '11, pages 8:1–8:8, New York, NY, USA, 2011. ACM.

[27] H. Wang and B. Wellman. Social connectivity in america: Changes in adult friendship network size from 2002 to 2007. *American Behavioral Scientist*, 53(8):1148–1169, 2010.

APPENDIX

A. INTERVIEW GUIDE AND QUESTIONS

At the beginning of the interview, we will not inform the interviewees the potential threats of accepting a strangers' friendship requests in Facebook. Our objectives is to collect interviewees' responses to investigate users' behaviors towards friendship requests sent from users and strangers in particular. Our sample includes active users on Facebook who logged in at least once a week.

Agenda:

1. Give an overview of the project: "The purpose of the study is to investigate the factors users employ when making a decision to befriend other users."
2. Introduce second interviewer and specify his role.

Part1:

1. General Questions:
   (a) What is your age?
   (b) What is your gender?
   (c) What is your highest level of education?
   (d) What is your major or occupation?
   (e) How long have you own a Facebook account?
   (f) How often do you use Facebook?
   (g) What is your first language?
2. The befriending behavior of users with strangers:
   (a) How many friends do you have on Facebook?
   (b) How often do you receive friend requests?
   (c) Have you ever accepted a friendship request from a stranger you do not know in real-life or have not met before online or offline?
   (d) What kind of factors do you rely on when you decide to accept a friendship request from a stranger? (For any factor users ask, we need to dig into more details by asking questions) (Gender, Friends, Mutual Friends, Profile, Picture, Wall show the activity in Facebook)
      - (The interviewee mentioned gender.) Will you accept a friendship request from a homosexual stranger or a heterosexual one?
      - (The interviewee mentioned friends.) How many friends does the stranger have that you will accept his/her friendship request?
      - (The interviewee mentioned mutual friends.) How many mutual friends does the stranger have that you will accept his/her friendship request?
      - (The interviewee mentioned profile.)
        i. Same/different hometown
        ii. Same/different schools
        iii. Same/different age
      - (The interviewee mentioned wall.)
      - Active/quiet person
3. Users' attitudes towards their privacy security:
   (a) Have you ever set your privacy setting? (If yes) How did you modify your privacy setting?
   (b) Have you assigned different privacy setting to your friends? (If yes) How did you modify your privacy setting for different friends?

(c) Have you had reported any security incident before in your online activities on Facebook, email, etc.?

(d) Have you realized that if you accept a friendship request from a stranger, he/she will have the access to your personal information? (If yes) What kind of information do you think will be exposed to the strangers?

(e) Do you mind your private data being exposed to the strangers? (If yes) What kind of information do you mind being accessed to the strangers?

4. Users' appeal of strangers:

   (a) How do you describe your connection with the stranger that you have accepted his/her friendship request?

   (b) Are you emotionally attached with the strangers?

   (c) At the very end, do mention that the request will be removed.

Debriefing happens here!

Part 2:

1. What would be your suggestion if you want to design the window for friendship requests?
2. Will you change your behavior towards friendship requests? (If participant had accepted the request)
3. Do you have anything else related to this study that you want to share with us?

B. SURVEY QUESTIONS

Thanks a lot for participating in this survey. In this survey, there are questions about your activities on Facebook. It will take you about 15 to 20 minutes to answer the questions. For the likert-scale questions, please choose one number from 1 to 5, where 1 means "strongly disagree" and 5 means "strongly agree".

1. What is your age?
   - 19 to 25
   - 26 to 30
   - 31 to 35
   - 36 to 40
   - 41 to 45
   - 46 to 50
   - 50 to 55
   - 56 to 60
   - 61 to 65
   - 61 and more
2. What is your gender?
   - Female
   - Male
3. What is your highest level of education completed?
   - High school
   - Undergraduate
   - M.Sc
   - PhD
   - Other:
4. What is your employment status?
   - Employed
   - Student
   - Retired
   - Unemployed
   - Other:
5. How long have you owned a Facebook account?
   - Less than a year
   - 1 to 2 years
   - 2 to 3 years
   - 3 to 4 years
   - 4 to 5 years
   - More than 6 years
6. How often do you login into Facebook?
   - Every hour
   - Several times a day
   - Once a day
   - Several times a week
   - Once a week
   - Several times a month
   - Once a month
   - I have my account de-activated
   - Other:
7. Please go to your Facebook profile. How many friends do you have on your Facebook profile?
   - Answer:
8. How often do you receive friendship request?

- Everyday
- At least once in 2-3 days
- At least once a week
- At least once a month
- At least once every 6 months
- At least once a year
- At least once in every two week
- Other:

9. Have you ever accepted a friendship request from somebody who you do not know in real life or online communities?
   - Yes
   - No
10. Check all groups that you would likely befriend on Facebook:
    - Parents
    - Siblings
    - Relatives
    - Close friends
    - Friends
    - Acquaintance
    - Colleagues
    - Other:
11. If I distinguish the person from the picture, I would accept the friendship request.
    - 1
    - 2
    - 3
    - 4
    - 5
12. I usually become friends with:
    - Only females
    - Only males
    - I do not care about the gender
13. Knowing the number of mututal friends is enough for me to accept a friendship request.
    - 1
    - 2
    - 3
    - 4
    - 5
14. If I have mutual friends with the person who sent me a friendship request, I would look at the closeness of those mutual friends to me in addition to just the number of mutual friends.
    - 1
    - 2
    - 3
    - 4
    - 5
15. If I know somebody in real world or online communities, I would accept her/his friendship request on Facebook.
    - 1
    - 2
    - 3
    - 4
    - 5
16. If I recognize someone's name, I would accept her/his friendship requests on Facebook.
    - 1
    - 2
    - 3
    - 4
    - 5
17. ( ) of my friends actively share content on Facebook (1: a few, 5: almost all)
    - 1 (a few)
    - 2
    - 3
    - 4
    - 5 (almost all)
18. I tend to accept friendship request from everybody, who was born in the s Iame city as I.
    - 1
    - 2
    - 3
    - 4
    - 5
19. I tend to accept friendship request from everybody, who lives in the same city as I do.
    - 1
    - 2
    - 3
    - 4
    - 5
20. I tend to accept friendship request from everybody, who have attended the same school/university as I do.
    - 1
    - 2
    - 3
    - 4
    - 5
21. Similarity in personal interests or hobbies is sufficient for me to accept friendship requests.
    - 1
    - 2
    - 3
    - 4
    - 5
22. I mostly accept friendship requests from people who share a lot of content on Facebook.
    - 1
    - 2
    - 3
    - 4
    - 5
23. Users who passively monitor others' posts on Facebook does'nt motivate me to post less content on Facebook.
    - 1

- 2
- 3
- 4
- 5

24. I limit my activities on Facebook because I know my friends are not interested in the content that I post.
    - 1
    - 2
    - 3
    - 4
    - 5
25. I don't tend to accept friendship requests sent from Facebook applications.
    - 1
    - 2
    - 3
    - 4
    - 5
26. I used to share more content since I felt more comfortable to share content with my Facebook friends.
    - 1
    - 2
    - 3
    - 4
    - 5
27. If my friends shared content irrelevant to me, I would remove them from my friends list.
    - 1
    - 2
    - 3
    - 4
    - 5
28. I don't accept a friendship request if I have just common interests or hobbies with the person who sent me friendship request.
    - 1
    - 2
    - 3
    - 4
    - 5
29. I would accept friendship requests sent from a Facebook application (for example a game) on behalf of others.
    - 1
    - 2
    - 3
    - 4
    - 5
30. Who is a Facebook user that you do not want to have a friendship connection with on Facebook?
    - Anybody who seems to be annoying (sending weird message, irrelevant post, etc.) regardless of being known in real life or not. 308
    - Anybody except people that are known to some extent
    - Anybody except for those that have strong connections in real life
31. How would you define different levels of access for Facebook friends?
    - Creating separate lists with different access levels
    - Using manual exemption feature for each shared content
    - I do not define different levels of access

To authorize or not authorize: helping users review access policies in organizations

Pooya Jaferian
University of British Columbia
Vancouver, Canada, V6T 1Z4
pooya@ece.ubc.ca

Hootan Rashtian
University of British Columbia
Vancouver, Canada, V6T 1Z4
rhootan@ece.ubc.ca

Konstantin Beznosov
University of British Columbia
Vancouver, Canada, V6T 1Z4
beznosov@ece.ubc.ca

ABSTRACT

This work addresses the problem of reviewing complex access policies in an organizational context using two studies. In the first study, we used semi-structured interviews to explore the access review activity and identify its challenges. The interviews revealed that access review involves challenges such as scale, technical complexity, the frequency of reviews, human errors, and exceptional cases. We also modeled access review in the activity theory framework. The model shows that access review requires an understanding of the activity context including information about the users, their job, their access rights, and the history of access policy. We then used activity theory guidelines to design a new user interface named AuthzMap. We conducted an exploratory user study with 340 participants to compare the use of AuthzMap with two existing commercial systems for access review. The results show that AuthzMap improved the efficiency of access review in 5 of the 7 tested scenarios, compared to the existing systems. AuthzMap also improved accuracy of actions in one of the 7 tasks, and only negatively affected accuracy in one of the tasks.

1. INTRODUCTION

Understanding and authoring access control policies has been known as a challenging problem [29, 33, 30]. But the focus of previous studies were on personal access control, where the data owner, policy maker, and policy implementer are the same person. This problem has not been extensively studied in organizational context. Bauer et al. [1] found that managing access control policies in organizations faces a unique set of challenges. In large organizations, those who make policies are different from those who implement these policies. Therefore, developing a shared understanding of policy between different stakeholders is challenging. In this paper, we explore and address this problem by proposing and evaluating AuthzMap, a new user interface for sense making and reviewing implemented access policies or, in short *access review*.

Access review is an important IT security activity in organizations, where the managers make the access policy and security administrators implement it. The managers are mandated by many security regulations (e.g., SOX [35], HIPAA [6]) to regularly review and validate the access privileges of users. However, Cser [10] suggests that access review for every 2,000 to 3,000 users consumes approximately one full-time-employee equivalent per year, and many organizations cannot even finish one access review process before a new campaign begins.

Recent security incidents that cost governments and organizations billions of dollars show the importance but yet lack the ability in reviewing users' access rights. For example, the US army soldier, Chelsea Manning, who leaked the US embassy cables was cleared to access classified resources when she was on training as an intelligence analyst. She then changed her job and location multiple times before going to Iraq. According to Swensen [34], if a superior reviewed Mannings' access and requested the revocation of unnecessary privileges, she would not have been able to leak the data.

The overarching goal of this paper is to investigate improvements technology support for access review. Towards this goal, we performed two studies. In the first study, we conducted 12 semi-structured interviews with security practitioners to understand how people make sense, and review access of users, and to identify the challenges in access review. We then designed a new interface, guided by activity theory guidelines by Kaptelinin and Nardi [19], to address the identified challenges. We named the proposed interface AuthzMap. We then conducted an online study with 340 participants to test if AuthzMap improves the usability over two of the existing interfaces.

Besides understanding access review activity and improving access review tools, this research has broader implications for the design of access management interfaces. Our results suggest that context plays a role in understanding the access privileges of an enterprise user. The context of a user-to-role assignment includes the user's current and past jobs, the history of the user-to-role assignment, other users' access privileges, and those who requested, approved, and implemented the access. Therefore, tools that manage user-to-role assignments should take into account the aforementioned information, and present them in a way that reflects the spatial layout and temporal organization of the context.

2. BACKGROUND

Organizations use many IT applications to run their business. Employees who use an application for their job are provided with a set of access privileges, and other employees should be prohibited from accessing the application. Therefore, applications provide a set of *permissions* that can be assigned to a *user* to control what the user is authorized to do. Sometimes, permissions are grouped into *roles* to simplify the provisioning process. As the number of users and applications grows, the management of users, roles, and permissions becomes challenging. Therefore, organizations are man-

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

dated by many security regulations (e.g., [35]) to frequently perform access reviews to make sure that users have the least set of privileges required for their job.

Next, we describe how access review is performed through an example. In an organization, security administrator, John, sends a request to manager Bob to review the access privileges of fifty employees who work in Bob's department. Bob is provided with the list of employees and their roles. He reviews the list of users one user at a time, looks at their roles, and verifies if the user-to-role assignments are valid. For example, Bob sees that Alice is assigned to 20 roles (R1, R2, ... R20). Bob needs to understand the meaning of the roles, what they authorize Alice to do, and if the authorizations are required for Alice to do her job. If an authorization is required, Bob *certifies* the assignment of Alice to the role. Otherwise, he *revokes* the assignment. If Bob cannot understand the meaning of a role, he may communicate with John or other managers to ask what privileges are associated with the role. This example shows that access review requires analysis, communication, and collaboration with other stakeholders.

3. RELATED WORK

There have been few studies related to access management in organizational context. Bauer et al. [1] performed a field study of access control practices in organizations. They suggest that the implemented access policy and the record of changes should be understandable and visible. Our findings confirm this, and our proposed interface improves understandability and visibility of access policy.

As opposed to access review, the problem of policy authoring has been previously studied. Brodie et al. [4] designed a privacy policy management workbench called SPARCLE to create policies in natural language. Although SPARCLE was successful in facilitating policy definition and management, it was not used or evaluated for the access review. Inglesant et al. [15] studied personal access control in Grid computing context. They showed that resource owners have difficulty expressing policies in RBAC and they prefer the use of natural language. Reeder et al. [29] proposed a new UI named "expandable grid" for understanding effective access policy in case of conflicting access rules. Expandable grid improves the understanding of access policy by end-users of commodity OSs, and their main goal is to address the issue with conflicting access rules that happen regularly in the Windows file system. The data from our interviews show that in enterprise environments, standard role-based access control without negative authorization rules is used. We also adopt the idea of expandable grid for use in an organizational context and use it in the design of AuthzMap. Smetters and Good [33] studied the use of policy authoring for personal documents. They found that users rarely change access policies, and tend to specify complex and error-prone policies. Our findings suggest that unlike access control for documents, the users' accesses change frequently in organizations. Vaniea et al.[36, 37] examined the effect of proximity of access management interface and the resources. They show that users detect errors better if controls are positioned near resources. Their proposed method was implemented and evaluated in the context of managing photo album privacy policy. In an organizational context, this proposal might not be possible, as resources do not have direct graphical representation, and the number of resources and permissions could be large. Beckerle and Martucci [2] identified six guidelines for designing usable access control rule sets, and showed that implementing those guidelines will help understandability of access policies. Their proposed solution can be used before presenting access control rule set in AuthzMap to reduce the complexity of policy.

Table 1: Interview participants' demographics

Code	Job title	Organization
P1	Security Manager	Insurance
P2, P8	Security Analyst	Insurance
P3, P7	Security Manager	Software
P4, P5	Security Administrator	Software
P6	Compliance Manager	Software
P9	Consultant	Health care
P10, P11	Consultant	Financial
P12	Consultant	Software

4. STUDY 1: UNDERSTANDING ACCESS REVIEW ACTIVITY

The initial goal of the interview study was to understand how organizations perform identity and access management, and the challenges they face. After initial analysis of interviews, we turned our attention to answering the following research questions: (1) Why organizations perform access review? (2) Who are the involved stakeholders? (3) Why access review is challenging? (4) How better decisions can be made during access review?

4.1 Methodology

We conducted 12 semi-structured interviews with security practitioners responsible for access management in large organizations. The list of interviewed participants, their roles, and their organization sectors are shown in Table 1. The scope of the interviews was various activities related to identity and access management (see Appendix A for the interview guide). The interviews were conducted by one or two interviewers in the workplace of the participant (8 interviews) or over the phone (3 interviews). The length of the interviews was between one and three hours. The interviews were audio-recorded, and transcribed.

We analyzed the interview data using grounded theory methodology [7]. We imported the transcripts of the interviews to a qualitative analysis software (Qualrus v2.1), and then coded them with open-coding technique with the codes emerging from the data. We then performed axial-coding by combining conceptually similar codes and identifying various themes across the data. At this step, we found that identity and access management involves several activities including access review. We also found different themes related to each activity including the goal, actors, artifacts, division of work, rules, and challenges. We identified access review as one of the most challenging activities. Therefore, we chose it as the core concept, and performed a round of selective coding to answer the research questions. We reached theoretical saturation [7] and stopped recruitment after recruiting 12 participants.

4.2 Results

In this section, we first provide a detailed description of access review activity using the activity theory framework, and then discuss the identified challenges.

We use the triangle model of activity proposed by Engeström [12] to lay out our description of access review (Figure 1). We will later refer to this formulation when we justify our design decisions.

The goal of the activity: Access review is an activity with the goal of verifying users' access rights to minimize the risk of unauthorized and unmanaged access and comply with regulatory legislations.

Subject: "Reviewer" is the main actor in the activity who performs access review. Our participants indicated that the following stakeholders act as reviewers:

Managers: Most of the participants indicated that managers review employees under their authority. P1 further described the role

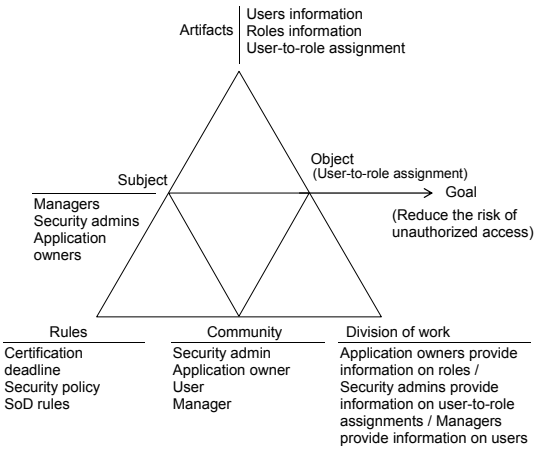


Figure 1: Overview of Access Review Activity

of the manager in access review: *"[Manager asks:] what access does Jim have? I'd like to review Jim's access because he's changing roles within my department, there's no official job posting but I'm doing a realignment and I would like to review Jim's access. So you need to do a specific report on Jim, which is to say here is the access profile that Jim has."*

Application owners: Two of our participants indicated that an application owner reviews the users who have access to the application, and certifies or revokes the users access privileges: *"Our team wrote some [scripts]. It goes out and it collects from these 80 or so applications, what the access lists are, what the rights are, it creates a report, we put it in a service desk ticket. Then it goes out to the [application owners] and they review it."* (P3)

Security administrators: P6 explained that his team is responsible for security compliance of a large enterprise application, and therefore he performs access reviews: *"We send a request to the manager that says Bob has changed from position A to position B. They are requesting position B roles. We are going to remove his position A roles. Do you agree with that?"*

Object: The object towards which the activity is performed is a user-to-role assignment. When managers or security admins perform access reviews, they review a set of roles assigned to a user (user access review). When application owners perform reviews, they review a set of users assigned to a role (application access review). We limit the scope of the AuthzMap to user access review. The same design techniques can be applied for building an interface for application access review.

Community and division of work: Access review involves security team members, employees, managers, and application owners. Involved stakeholders divide the work as follows: A member of the security team requests review of users' access rights. The reviewer (a manager in most cases) receives the request. He goes through the list of users, selects a user, and identifies the user's roles. For each user-to-role assignment, he chooses to certify or revoke the assignment. The reviewer might contact the application owners, the user, or the security team when he is unable to determine the correct action.

Rules and constraints: Different rules and constraints impact access review. (1) The security policy of the organization determines the validity of a user-to-role assignment. For example, P9 explained that in health care, they follow an optimistic security paradigm [28] and allow more access than usual so the physicians can access patients' files in emergency cases: *"So the whole access model in health care tends to be, you let people do what they need to do to get the job done."* (2) Static separation of duties (SoD) rules determine if a user can be assigned to two or more specific roles at the same time. (3) The review deadline set by security team constrains the time window of the review.

Artifacts: Reviewers use three artifacts during access review: (1) User's information, which include the identity related information, the job title, and other attributes like the phone number, email, department, etc. (2) User-to-role assignment information, which include who requested, who approved, and who implemented the assignment, when and why the user was assigned to the role, and who previously reviewed the assignment. (3) Roles' information, which include the role's name, description, the owner, and the permissions assigned to the role.

4.3 Challenges in access review

Our interviewees indicated that access review is a challenging activity. We classified these challenges into 5 categories:

Scale: Access review can involve large number of users, roles, and permissions. P6 explained that just one of the large applications in his organization has 16,000 users, up to 115 roles per user, and up to 407 permissions per role. He also indicated that reviewers have to review up to 200 users in a review activity. While these numbers vary from application to application, and from organization to organization, they show the magnitude of data that a reviewer needs to deal with.

Lack of knowledge: When managers act as reviewers, they do not have the expertise to understand the meaning of roles and permissions. P2 illustrated this problem in detail: *"we send these godawful long reports to the new manager hiring the employee is going into, saying "let us know which access this person needs to keep and what they need to remove." And a lot of it's, you know, cryptic RACF information and stuff they just have no idea what they're even reading so they either take their best guess and say, ok, then maybe this sounds kind of like something they might need. Or they just say they need it all."*

Frequency: While reviewing access is not the main job of managers, they are frequently asked to perform this activity. For example, P3 explained why they perform quarterly access reviews: *"... Once a quarter! We do quarterly access reviews. [...] Once a year is never good for any control because if you fail, you fail; at least twice a year you have a chance to remediate."* Additionally, P3 talked about ad-hoc access reviews: *"Every day, [access management software] looks at [every] person who has access and says has the person changed in any way. Did they move departments, did they move to geographical locations - if so it triggers an event which puts a ticket into the service desk system, sends a note to the Access Reviewers and says you need to review this ..."*

Human Errors: P3 described why human errors are common during reviews: *"So the policies of the company states that the business is responsible for the access. So the ultimate decision maker is the business. However they failed because it's a human process right? It's eyeballing [and] sometimes the lists are large."* Such errors would be costly for organizations, both in terms of leading to data breaches, and failing compliance reviews.

Exceptional Cases: In organizations, the validity of user-to-role assignments cannot be determined accurately only by knowing the user's job function. Users might need to fill in another employee's role for a period of time, or they might need temporarily access certain resources when they are on training. P6 explained a case where they thought they should remove existing access from a user because he asked for new access. They later realized the user is on training and still has his old job: *"The manager says no, he is training this person, as replacement, for three months."*

5. AUTHZMAP DESIGN GOALS

To design a new access review tool, we followed a design approach proposed by Kaptelinin and Nardi [19]. In this section, we present three main design goals. For each goal, we first present the theoretical support, and then we use the field study data to describe how we apply theory to the design of an interface for access review.

Flexible support for review actions: The goal of access review is verifying access privileges. This goal can be broken down to lower level subgoals, and actions to satisfy those subgoals. These actions can include: viewing list of users and identifying them, identifying users' job function, checking the list of users' roles, and certifying or de-certifying user-to-role assignments. To address the *Scale* challenge, a tool can help users perform the above actions more efficiently. This can be achieved by more flexible search and filtering mechanisms to view and identify users, and applying decisions in batch. Technology should also support alternative ways to attain an activity goal [20]. To achieve this, we present information at different levels of abstraction. The user can choose the right level of detail, based on his knowledge and understanding of the access policy. For example, a user with the knowledge of the access policy can use more abstract view, but a user who needs more information can use detail view. This approach can address the *lack of technical knowledge* challenge.

Visibility of context: Activity theory emphasizes that tools and artifacts used during an activity are part of the context, and the technology should facilitate access to those artifacts, integrate them with each other, and present them in a way that reflects the spatial layout and temporal organization of the context. The context of an access review activity includes users, roles, and user-to-role assignments. In addition, the following artifacts are part of the context and can be used for making access review decisions:

(1) Job changes: Our participants indicated that when users change their job or move between departments, their access changes. For example, P6 explains why job changes can be an important contextual information for access review: *"Now what happens is that we have a report that runs every single day and it tells me [if] people transfer [to another department] or change [their job]. [For example,] she gets a promotion. She went from warehouse manager to public relations manager. She will request something. I need a public relations manager role. My team goes automatically: 'why? That's not what you are. You are warehouse. No, I got a promotion, I'm this. Okay, we'll give you these three but you are losing those three."'* Providing job changes help reviewers better understand how and why the access privileges of users change, and therefore, address the *lack of knowledge* challenge.

(2) Other users' access: During access review, reviewers may need to review many users instead of one. These users have certain roles in common (e.g., basic access to the Internet, email, Sharepoint). For example, P1 explained that users who are doing the same job usually have similar access: *"...a manager who hired a new employee [and] who knew that you had the access that you needed to do the job for him or her would say, 'Oh, make this new employee's access just like yours.' And so then an employee would then inherit privileges based on the success of a previous employee in terms of doing that job."* Therefore, comparing access privileges of a user known to reviewer to that of an unknown users will facilitate sense making. This will address *lack of knowledge*, and *scale* challenges and reduces *human errors*.

(3) Previous reviews: The reviewer can employ the past review decisions and replicate them in his review. Replication is particularly useful if none of the user's attributes has been changed since the last review. Having access to and using past reviews can address *frequency*, and *scale* challenges, and reduces *human errors*.

(4) Other users involved in the activity: The process of provisioning users with access privileges is a collaborative activity between different stakeholders. Therefore, the interface should show who requested the access, who approved the request, and who executed the provisioning of access. (P12) explained that such information will help reviewer make an informed decision: *"So again, you think of the attestation process or even at any moment in time on a view user, we always talk about helping somebody make informed choice. So if I'm evaluating the correctness of an SAP account and I can look at when it was requested, who reviewed it, who approved it, when your last login time was, I can serve to make a pretty informed choice about why you have this or its level of appropriateness."* This can address *lack of knowledge* of why a user has certain access privileges.

(5) Policy violations: Our previous survey [16] shows that SoD violations are the most important violation to be detected during access review. Therefore, they should be highlighted on the interface. This can address *scale*, and *lack of knowledge*.

Make history visible: According to [19], analysis of the history of an activity can reveal the main factors influencing the development of the activity. Furthermore, Hollan et al. [14] studied experts working in complex environments, and suggested historical information can be incorporated in cognitively important processes. For access review activity, historical information can help reviewers understand how the policy has evolved over time, and therefore make better decisions in uncertain scenarios. This would address the challenges of *scale*, and *exceptional cases*.

To incorporate history in the interface, we first identified which of the three access review artifacts (users, roles, and user-to-role assignments) carry historical information. Interview data revealed that users, and user-to-role assignments (unlike roles) change over time, and therefore, have historical information. For example, P6 explained that employees frequently change their job, but roles should be designed in a way that are not impacted by such changes: *"[Employees' position] changes a lot when you start going through economic churns. So when you are laying-off 50 people at a time, 100 people another time, or department consolidations. I can tell you I've been in this role for two and a half years and I've seen five department consolidations in finance alone. So when you see all those changes happening, those composite roles hurt you. Because then you have to keep generating them over and over again."* Also when we asked P4 about how frequently they make changes to the roles, she responded: *"We don't. I wouldn't say never - very rarely. If we were to add a new region, which I don't think there are any left to be added at this point."* Therefore, AuthzMap visualizes the history of users' job changes, and the history of user-to-role assignments, and correlates them with each other. Showing the history can addresses *frequency* challenge by showing previous decisions, and help with understanding of *exceptional cases*.

Knowledge sharing: According to Kaptelinin and Nardi [19], technology should help in problem articulation and seeking help from colleagues. The interview participants indicated that reviewers hardly understand the meaning of the roles and access privileges. Therefore, our participants used the following strategies to mitigate the lack of knowledge:

(1) P6 talked about translation of technical terms to business related terms to help reviewers understand the meaning of roles: *"... and we get this huge profile - here's all the access the user has. We then have to translate that into more of an English format for the individual."*

(2) P7 described the use of communication channels to get help with certification decisions: *"The security coordinators take it to the [application owner] and explain what the risks are. They're the*

ones who do a kind of mini risk assessment say: OK, such and such business unit wants access to this data for such and such reason."

Therefore, one of the design goals in the proposed interface was to provide knowledge of each access privilege for the reviewers in the form of a description, and list of permissions (in case of using roles). Moreover, communication channels should be available in the interface to get help from other users with the knowledge of roles and permissions. Knowledge sharing would address the challenges of *lack of knowledge*, and can help with *exceptional cases*.

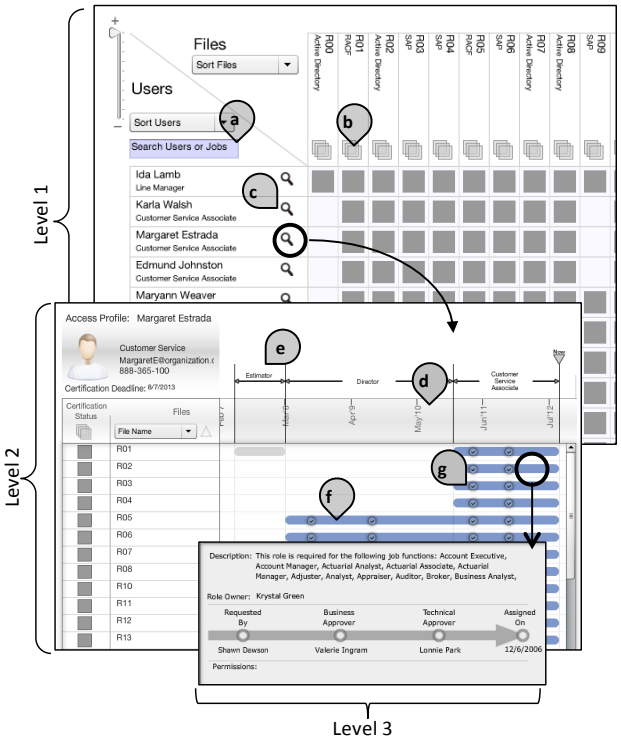


Figure 2: The three levels of the AuthzMap interface. The reviewer is presented with Level 1 of the interface. He can go into Levels 2 and 3 for making further sense of the accesses of the users.

5.1 AuthzMap Interface Design

To realize the goals discussed in the previous section, we designed a new interface and named it AuthzMap. We first built a low-fidelity prototype in Microsoft Visio, and improved it over multiple rounds of internal feedback. We then designed a medium-fidelity prototype in Adobe Flash, and refined it by getting feedback from external usable security researchers, as well as our industrial partner in this project. Finally, we built a high fidelity prototype in Adobe Flash. It loads access control related data through XML files and allows the user to perform access review tasks. We depict the AuthzMap in Figure 2, and with more details in Appendix B.

The AuthzMap uses three levels of abstraction to integrate different contextual artifacts discussed in the previous section. Level 1 shows users and roles in a grid that provides an overview of the overall review activity. The spatial layout of the interface was based on Lampson access matrix [22] model and inspired by the design of Expandable Grid [29]. Users are sorted from top to bottom, based on the number of privileges they have. This allows reviewers to quickly identify users who have large number of privileges. Reviewer can use the sorting and filtering functionality (Figure 2a) to group and compare users with similar job titles, or roles that serve access to similar applications. AuthzMap provides batch certify accelerators (Figure 2b) to certify a role for all users. Reviewer can obtain the detailed access profile of a user using the second level of the interface (Figure 2c).

In Level 2, we integrated contextual information related to the user-to-role assignments, the job changes of the user, and previous reviews. This level also uses a timeline metaphor (Figure 2d) to show the temporal relationship between the job changes (Figure 2e), roles (Figure 2f), and previous reviews (Figure 2g). The reviewer can re-arrange the roles based on the role name, active roles, and the time the role is assigned to the user.

If the reviewer needs more details on one particular user-to-role assignment, he can click on the role bar to go to Level 3 of the interface, which shows the description of the role, the role owner, the permissions assigned to the role, and the workflow through which the user obtained the role. Level 3 allows the reviewer to learn about the meaning of the role. If the reviewer cannot articulate the meaning or the impact of the role using this information, he can use communication channels to seek help by clicking on the name of each stakeholder (e.g., owner, requester, approver, and implementer).

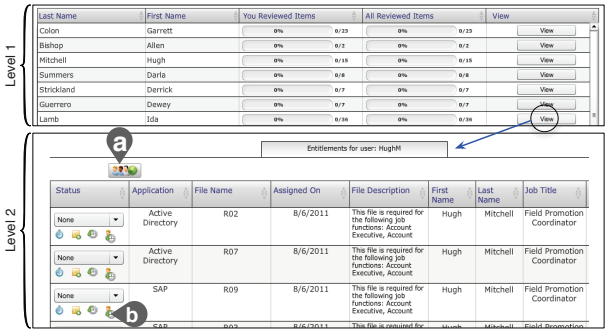


Figure 3: A screenshot of the List interface. Reviewer identifies the user and clicks on the View button. Reviewer is presented with the second level of the interface that includes the list of user's access privileges. The icon marked as (a) allows batch actions on privileges, and the four small icons (marked as b) do the following (from left to right): sets the access expiry time, writes notes for each privilege, shows history of actions on each privilege, and shows history of rejections for each privilege.

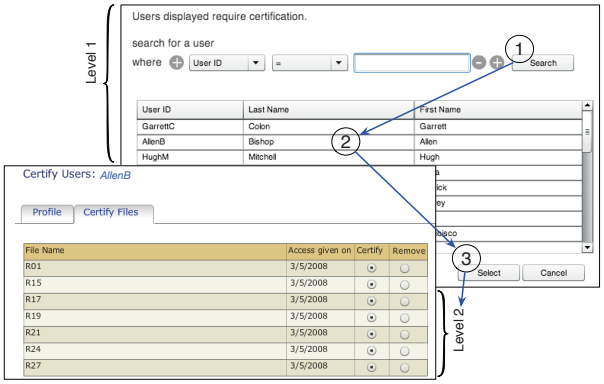


Figure 4: A screenshot of the Search interface. (1) Reviewer searches for a user. (2) Selects the users. (3) Clicks on the Select button and certifies or revokes access privileges in Level 2.

6. STUDY 2: EVALUATION OF AUTHZMAP

The goal of our evaluation was to test if AuthzMap is more usable than the two existing systems. Nielsen defines usability by five quality components [26]: (1) Learnability, (2) Efficiency, (3) Memorability, (4) Errors, and (5) Satisfaction. In this study, we identify efficiency and errors as two main usability goals of the interface, as they are directly related to the challenges described in Section 4.3. At the end of the study, we also collect data about subjective satisfaction of the participants.

6.1 Evaluation Methodology

To evaluate three interfaces, we designed a between-subjects study with 3 conditions (one condition per interface). We asked participants of each condition to perform seven tasks. For each task, the interface was the independent variable, and we measured the following dependent variables: (1) efficiency, by recording time to completion (TTC), and (2) accuracy, by recording correctness of the critical components of the task.

6.1.1 Evaluated Interfaces

We compared the AuthzMap interface to two other interfaces, named Search (Figure 4) and List (Figure 3). The detailed description of each interface is provided in Appendix B. The List interface is known as one of the two access review market leaders, and Search as one of the two the strong performers [10]. We choose not to reveal the actual names of Search and List interfaces as the purpose of the study is not to critique a particular commercial system, but rather compare three different approaches in the design of access review interfaces. The Search interface does not reveal the context at all. A reviewer can search for users, select users one-by-one, and review the user-to-role assignments. The List interface reveals certain contextual information such as the progress of reviewing individual users, the history of previous reviews, and information about the individual users (such as their job and department) and the roles (such as the date the user is assigned to a role, or the description of the role). But these contextual information are not correlated with each other or immediately accessible to the user. We chose to build a prototype of Search and List interfaces over using their full versions for two reasons. First, we wanted the three interfaces to be at the same level of granularity. Second, we did not have access to the installable version of the List interface. Third, prototyping allowed us to instrumentalize the interfaces for the user study.

6.1.2 Participants

We used Amazon Mechanical Turk (MTurk) for recruitment, and paid each participant $2. MTurk has been used as a user study platform for HCI [21] and usable security research [38]. Participants were asked to play the role of managers responsible for access review. Because we did not specifically recruit managers, we used an approach similar to the one by Convertino et al. [9], to provide participants with the beliefs and knowledge of managers. Using our interview data, we first determined managers' level of computer security, review tool, and organizational knowledge. Interviews showed that managers do not have an extensive computer security knowledge, but they understand the concept of access review, and they know the steps for performing it. In addition, managers are trained on using the access review tool (i.e., they are not the first time users of a novel tool). We also assume they are not daily users of the tool (they use it four to two times a year or on an ad-hoc basis). To help participants have similar level of knowledge, we trained them on the basics of access review, and the use of tool to perform reviews (see Section 6.1.3 for the details of our training procedure). We further allowed them to explore the tool and familiarize themselves with it.

6.1.3 Training Material

We designed training material to ensure participants understood the concept of access review, and could apply that understanding using the system. The participants were given a brief training on access control and access review. We followed the recommendations from previous research on designing training materials:

Brief, up to the tasks: Users will learn tools faster when the training focuses on performing the task rather than understanding the rationale behind the task [5]. We avoided training users on details of role-based access control, and concepts such as roles, and entitlements. Instead of using the notion of roles, entitlements, or access privileges, we used the notion of access to files. Previous research shows that participants can understand the meaning of file access control, and they are able to comprehend file access control policies [2, 29, 30].

Use of examples: We used examples throughout the training to explain the access review concepts. We also provided instances of how the interface can be used in interpretation of users' accesses.

Use of text-based material: Online participants can do better with short textual instructions, rather than videos, or demos [13] as it gives participants the opportunity to easily revisit the training during the study.

Use of multi-staged training To avoid overloading participants with training material, and to help them start working on tasks as soon as possible [5], we only taught them the basic access review concepts during the training. Task specific topics such as separation of duties (SoD) violations, privilege accumulation, etc. were taught as parts of the scenarios.

After the training, participants were asked to complete a test to check if they have the required knowledge to do the tasks. We tested the understanding of access control and access review using six multiple choice questions. Multiple choice questions are a reliable and objective way to assess the outcome of the learning, while the answers can be checked automatically [8]. We used standard techniques for designing multiple choice questions [8], and piloted them to ensure their effectiveness.

6.1.4 Study Material

Actual users of access review tools also posses the organizational and contextual knowledge that our participants lacked. For example, a manager may have an understanding of the consequences of having access to a resource, or awareness of the access privileges for doing certain job. Such knowledge is context dependent, that is, we cannot have a clear assumption that a manager always has or lacks such understanding. In the study tasks, we simulated both situations where reviewer has or does not have contextual knowledge and provided participants with documents and material as external knowledge sources (similar to [9]).

We presented participants with three documents: *file catalog*, *application catalog*, and *SoD catalog*. The file catalog showed the list of files that each job function was allowed to access. The interview participants talked about entitlement catalogs, which we changed to the file catalog for the purpose of this study. According to P7: *"One of the things we have been doing is also building a catalog of access requests that people can make [based on their job]."* The

application catalog listed all the applications and their files (entitlements). According to P3, they kept the track of this information in a knowledge base: *"our access procedures state that every application that has any level of criticality is supposed to have a published knowledge-based document in our service desk knowledge base that defines what the application is ..."* The SoD catalog showed pairs of files that caused SoD violations. P12 said they document these rules: *"And again whether they be SOD policies that say you can't have A if you have B or what we call 'restricted access' policies that say you can't have entitlement X if you are not cost center Y or division X or whatever the rule is, the ability to define that rule it lives with the entitlement in the resource catalog."* (P12)

Norman [27] describes that people can rely on *knowledge in the world*, *knowledge in the head* or a combination of both in their activities. To determine the validity of a user's access, reviewers may completely rely on the above documents (knowledge in the world), they may completely rely on their own knowledge (knowledge in the head), or they may use a combination of both. The lab study participants did not have the knowledge in the head of the hypothetical organization. Therefore, all the required knowledge for performing the tasks was included as knowledge in the world in the form of provided materials during tasks.

6.1.5 Study Tasks

After completing the training and the training test, participants were asked to perform seven tasks. We aimed to design tasks with three characteristics [24]: (1) Realistic; (2) Actionable; (3) Avoid Clues or Steps. In order to achieve realism, we designed the tasks based on interview data and a survey we previously performed [16]. Tasks #2 and #3 simulate conditions where the manager knows which access privileges are appropriate for users, and only needs to identify users, and certify or revoke the privileges. Tasks #4 to #6 simulate scenarios where the manager tries to detect access privileges with high risk. To further understand what type of access privileges are risky, we conducted a survey [16] and asked participants to rate the risk associated with certain types of access privileges. We chose to use the top three, which were SoD violations, accumulated privileges, and access privileges to critical applications, and used them to design tasks #4 to #6. The task #7 was a combination of previous scenarios to simulate a more uncertain and complex situation.

Training Task: The goal of this task was to familiarize participants with the interface. As we described in Section 6.1.2, managers will be familiar with their access review tool. This task gave participants an opportunity to perform a guided exploration of the interface, and understand how they can find pieces of information required in the upcoming study tasks. Participants were given the following scenario: "You are asked to identify the following information about "Clay Warren" : (1) his current job title, (2) list of files he has access to, (3) his previous job title, (4) the date of the last access review performed on the user." They were expected to select the correct answer to questions #1, #3, and #4 from seven possible options (including "I do not know"). They should also type the answer to question #2 in a text box.

Common Review: (P1) explained that a common access review scenario is when a manager reviews one user: *"[Manager says:] what access does Jim have? I'd like to review Jim's access because he's changing roles within my department, there's no official job posting but I'm doing a realignment and I would like to review Jim's access."* Therefore, participants were given the following task: "You are asked to review the files *Timothy Larson* has access to. Check the user's access to files, certify the access to those files the user requires to perform his job and revoke those he does not require. Feel free to use the *File Catalog* in the top menu to find the list of files required for performing each job." In this scenario we made an assumption that the manager can determine the correct set of access for users. This is simulated by providing participants with access to a *File Catalog* that shows what files are required for performing each job. Participants are expected to revoke access to two files that are not necessary for Timothy Larson's job.

User comparison: P2 explained that similarity between users with the same job is used to detect excessive and unncessary access: *"if you've got a group of 15 case managers and you bring them into the system, it'll say: ok, 12 out the 15 have 80% of access in common and these two people only have 20%. [...] oh this person has access they should not have, that has been carried over from somewhere else."* To simulate this scenario, participants were given the following task: "In this task you need to certify the access of three users. The certification is only limited to employees with *Loss Control Consultant/Specialist* job function. Identify such users, certify the files that users require to perform their job and revoke the access to the files they do not need. The catalog of jobs, and the required files to perform each job will be provided." In this scenario, we made an assumption that the manager can determine the set of access privileges required for the job and therefore provided participants with *file catalog*. In this task, there were three users with the "Loss Control Consultant/Specialist" job, and one of them had an unnecessary access to a file. Participants were expected to revoke the access to that file.

Privilege Accumulation: Many of our interview participants discussed the privilege accumulation problem in large companies. For example, P6 explained: *"I was warehouse worker, I became public relations. They would request the public relations roles, nobody would take away the other ones and you would wind up with somebody having 50 roles."* Therefore, this task evaluated the interface in finding and resolving accumulated privileges. We gave the participants the following scenario: "Assume you do not know the list of files required for performing each job. In this case, you need to evaluate each user's access to files based on the following rule: *If the user changes job, he should not keep any access from his previous job. Any access that is kept from a previous job should be revoked.* Please review the following users, and revoke invalid accesses according to the above rule: (1) Derrick Strickland, (2) Lynda Robertson." The two target users had two and one permission accumulated from their past job, and participants were expected to revoke those permissions.

SoD Violation Detection: P6 described SoD violations as one of the highest access related risks. He described a case that someone is moving from accounts receivable (AR) to accounts payable (AP), and access to AR and AP systems causes SoD violations: *"So you are going from - you are the AP person, you are going to AR and your AP person needs to be trained [by you] – your replacement. Then we don't like it and it becomes very problematic and we usually want lots and lots of controls if you want the person to have the access."* Therefore, the goal of this task was to evaluate the proposed interface in the detection of SoD violations. We gave the participants the following scenario: "Sometimes a user should not have access to two specific files at the same time. For example, a user can have access to either file A or B but not both, at the same time. This rule is called *Separation of Duties (SoD)*, and having access to those files at the same time is called an *SoD violation*. In this scenario, you are asked to review the files of two users, and detect and eliminate SoD violations. To do so, you should first identify the two files that cause SoD violations, and remove access to one of the files to eliminate the violation. Please check the following

users for SoD violations: (1) Ida Lamb, (2) Maryann Weaver." In this task, each of the users had access to two files that caused SoD violation, and participants were expected to revoke access to one of the files causing SoD violation.

Application Review: P3 noted that they sometimes prioritize the access review according to applications. Critical applications are reviewed first, and in some instances non-critical applications are excluded from the review: *"They run a process which goes out to a subset of all those applications - the ones that we call critical which is SOX applications plus other [...] It goes out and it collects from these 80 or so applications what the access lists are, what the right are, it creates a report, we put it in a service desk ticket. Then it goes out to the [reviewers] and they review it."* In this task, we evaluated interfaces for application specific reviews. We gave the participants the following scenario: "The company uses four applications for running the business: *Active Directory*, *Great Plains*, *RACF*, and *SAP*. Each of these applications uses a subset of the available files. You are asked to review the following users, and revoke access to the files related to the *Great Plains* application: Edmund Johnston, Nelson Murphy, Jane Hoffman, Olive Morris." The four users in the scenario had access to 27, 21, 15, 2 files respectively, out of which 7, 5, 3, and 0 files were related to "Great Plains". Participants were expected to revoke access related to the *Great Plains* application.

Comprehension Task: In the previous tasks, we evaluated interfaces for specific scenarios, and told participants to look for a specific situation. In reality, reviewers may deal with a combination of various scenarios and need to integrate various cues to make decisions. This task aimed to evaluate the interface for situations where reviewer needs to evaluate the risk of particular access in the presence or absence of various indicators of risk and safety. We gave the participants the following scenario: "You are provided with a list of users and their accesses, and you are asked to determine how risky access to each file is. Use the knowledge you gained during the previous tasks to determine the risk associated with each file: (1) Francisco Lee, Director, R06; (2) Marcella Owens, Claims Manager, R02; (3) Margaret Estrada, Customer Service Associate, R11; (4) Alyssa Jacobs, Customer Service Manager, R09"

For each of the four user/file pairs, participants were asked to rate the risk associated with the user having access to the file using a five point likert scale (1= Very Safe, 5 = Very Risky). The order of the four likert scale questions was randomized. Four user/file pairs had different levels of risk associated with them: (1) *Marcella Owens, R02*: Access to R02 caused a separation of duties violation with R44. We expected the participants to rate the risk at 5 (High risk). (2) *Francisco Lee, R06*: Access to R06 was given to the user during his previous job. Also there was no previous review of the user's access. On the other hand, there was another user with the "Director" job title who also had access to R06. We expected participants to rate the risk at 2, 3, or 4, as this access was associated with both indicators of risk and safety. (3) *Margaret Estrada, R11*: Access to R11 was given to the user as part of her current job, the access was certified twice during past reviews, and the two other users with the same job as Margaret had the same access. We expected participants to rate the risk at 1 (High safety). (4) *Alyssa Jacobs, R09:* Access to R09 was revoked from the user during a previous review, but the user gained access again after a while. Furthermore, other users with the same job did not have access to R09. We expect participants to rate the risk at 5 (High risk).

6.2 Analysis

The goal of our analysis is to compare the three tested interfaces in terms of efficiency, and accuracy.

Efficiency: We used time-to-completion (TTC) as a metric for efficiency. To capture TTC, we automatically logged the time users spent between starting and finishing each task. Then for each task, we tested the following null hypothesis: H_o: There is no difference between the median time to completion when using any of the three interfaces. H_1: There is a difference between time to completions. We used Kruskal-Wallis test, which is a non-parametric alternative to ANOVA, since we found that the time to completion was not normally distributed, and we could not normalize the distribution using transformation. Whenever we rejected the null hypothesis, we used pairwise Wilcox test with Bonferroni adjustment to test the following three null hypotheses: (A=L) There is no difference betweeen AuthzMap and List. (A=S) There is no difference between AuthzMap and Search. (L=S) There is no difference between List and Search. For each test, we report the p value and the effect size (r). We also discuss the practical significance of the difference between AuthzMap and the other interfaces by showing the percentage of improvement or declination of median TTC over the other interfaces.

Accuracy: We identified those critical components of each task in which participants can commit dangerous errors. An error is dangerous if it puts the system in insecure state (i.e., leaves user with excessive privileges). For each critical component, we calculated the total number of participants who did and did not commit the error. Then we tested the following null hypotheses: (1) (A=L) There is no difference between the correctness of answers of AuthzMap and List participants. (2) (A=S) There is no difference between the correctness of answers of AuthzMap and Search participants. We used two-sided Fisher's exact test with Bonferroni adjustment to test the above hypotheses.

6.3 Results

In this section, we present the results of our data analysis. First, we provide a summary of participants' demographics and experience. Then we present the findings of the study. In this section, we use abbreviated condition names when presenting the results (A = AuthzMap, L = List, S = Search). Table 2 shows the number of participants who consented to the study, attempted the study, finished the study (received a return code for compensation), and those who provided valid results. If participants clicked on the consent form, we counted them as a consented participant. If a participant at least started the background questionnaire, we counted them as an attempted participant. If a participant completed all of the stages of the study, we counted them as a finished participant. Some of the finished participants skimmed through the study (our system recorded their time to completion for certain tasks at 0 seconds), or intentionally or unintentionally bypassed our system in order to get to the finish page without completing all of the tasks. We eliminated these participants, and reduced the pool of participants to a set of valid participants. We made use of data from 430 valid participants in this section.

We also tested the following null hypothesis: H_o: The validity of participants is independent from the interface they were using. To test this hypothesis, we divided the attempted participants in each condition into two groups: those who were valid participants, and those who were not valid participants. Our chi-square test revealed that the validity of the participants depends on the interface ($\chi^2(2, N = 1030) = 20.424, p = 3.7e-05, Cramer'sV = 0.141$). However the effect size is small.

We show the total time needed to complete the entire study for the valid participants in Figure 5. We tested the following null hypothesis for the time to completion of the study: H_o: The choice of the interface does not impact the total time needed for comple-

Table 2: Classification of participants according to their progress in the study

	A	L	S	Total
Consented	355	355	354	1064
Started	341	341	350	1032
Finished	190	156	151	497
Valid	174	135	121	430

tion of the study. A Kruskal-Wallis test revealed a significant effect of interface on the time to completion of the study ($\chi^2(2) = 48.033, p = 3.7e-11$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed significant differences between AuthzMap and List ($p = 4.8e-10, r = 0.31$) and between AuthzMap and Search ($p = 6.4e-07, r = 0.25$).

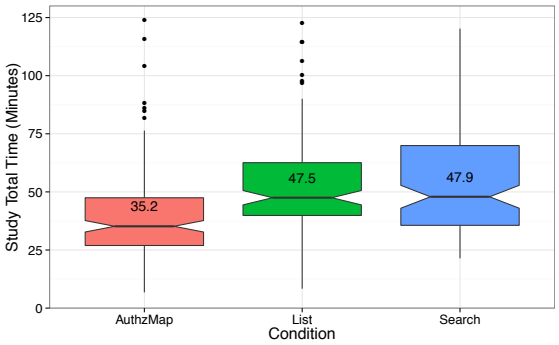


Figure 5: Total time needed to complete the study for participants in each condition

6.3.1 Participants Demographics

In the beginning of the study participants were asked to answer the background questionnaire. We show the overview of the participants' responses in Tables 3.

Table 3: Participants Demographics

		A	L	S	Total
Gender	Female	46.6%	52.6%	56.2%	51.2%
	Male	53.4%	47.4%	43.8%	48.8%
Education	Less than High School	2.3%	0.7%	0.8%	1.4%
	High School, diploma	8%	12.6%	10.7%	10.2%
	University/College Deg.	86.8%	85.9%	86.8%	86.5%
	Professional Deg.	2.9%	0.7%	1.7%	1.9%
Age	18-24 years old	30.5%	25.9%	42.1%	32.3%
	25-34 years old	43.7%	54.8%	39.7%	46.0%
	35-44 years old	15.5%	11.9%	10.7%	13.0%
	45-54 years old	6.3%	5.9%	5.8%	6.0%
	55-64 years old	3.4%	1.5%	1.7%	2.3%
	65-74 years old	0.6%	0%	0%	0.2%

6.3.2 Training

Participants were asked to complete the post-training test before proceeding to the study tasks. We summarized the number of attempts to complete the test in Figure 6. The results showed that nearly half of the participants in each condition could pass the test in the first attempt.

6.3.3 Per Task Results

In this section, we compare three conditions per task. Table 4 shows the median time to completion of individual tasks. The result of Kruskal-Wallis test for each task showed a statistically significant difference between three conditions. Therefore, we only show the result of three pairwise comparisons between conditions in Table 4.

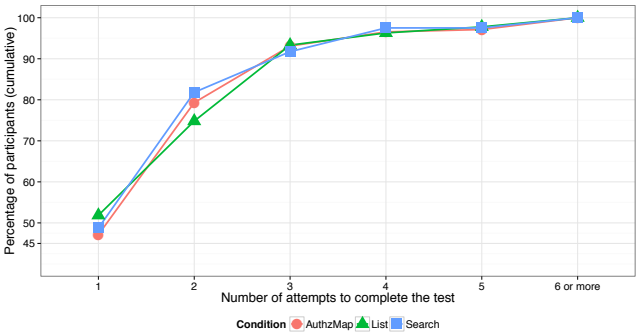


Figure 6: Number of attempts in completion of training test

Training Task: Table 4 shows that AuthzMap improved efficiency over the two other interfaces, although the effect size was medium. In terms of practical significance, AuthzMap reduced time to completion by about 20% compared to List, and by 25% compared to Search. Table 5 shows the results of the accuracy analysis. Fewer participants in AuthzMap condition committed errors in identifying the last job function of the user, compared to List condition, and in identifying the date of last access review, compared to Search condition.

Table 5: Comparing the correctness of participants' responses to the training task

	A	L	S	A=L	A=S
Job Title	97.1%	97%	98.3%	1	1
List of files	87.4%	80%	90.1%	0.443	1.000
Last Job	87.4%	54.1%	81%	<0.05	0.738
Last Review	75.9%	66.7%	35.5%	0.394	<0.05

Common Review Task: Table 4 indicates that for reviewing a single user, while the reviewer knows the files the user should have access to, Search is the fastest interface. Yet, looking at the effect size reveals that the size of the difference between AuthzMap and Search is small. In other words, AuthzMap reduces the median time-to-completion by approximately 17%, compared to list, but increases time to completion by approximately 25%, compared to Search. In this task participants could commit two dangerous errors (i.e., not revoking invalid access), and we show the proportion of participants who correctly revoked such access in Table 6. Table 6 shows that we rejected all four accuracy hypotheses, and shows that participants in AuthzMap condition had more errors than the two other conditions.

Table 6: Comparing the correctness of participants' choices in common review task.

	A	L	S	A=L	A=S
Revoked R19	70.7%	86.7%	88.4%	<0.01	<0.01
Revoked R10	70.1%	87.4%	87.6%	<0.01	<0.01

User Comparison Task: Table 4 shows that AuthzMap improves efficiency over the two other tasks. In terms of practical significance, AuthzMap decreased the time to completion by about 105%, compared to List, and by about 78%, compared to Search. The accuracy analysis (Table 7) did not reject any of the accuracy null hypotheses.

Table 7: Comparing the correctness of participants' choices in user comparison task.

	A	L	S	A=L	A=S
Revoked R13	84.5%	88.9%	86.8%	0.632	1.000

Privilege Accumulation Task: Table 4 shows that AuthzMap improves efficiency over the two other tasks. In terms of practical significance, AuthzMap improved time to completion by about 186%, compared to List, and by about 112%, compared to Search. Table 8 shows the result of accuracy tests. We rejected three of the null hypothesis for comparing AuthzMap and List, but we did not reject any of the hypotheses for comparing AuthzMap and Search.

Table 8: Comparing the correctness of participants' choices in privilege accumulation task.

	A	L	S	A=L	A=S
R06, LyndaR	86.8%	68.9%	79.3%	<0.05	0.652
R03, DerrickS	88.5%	71.9%	80.2%	<0.05	0.4
R12, DerrickS	86.2%	71.1%	81.8%	<0.05	1

SoD Violation Detection Task: Table 4 shows that AuthzMap improves the efficiency of detecting SoD violations. In terms of practical significance, AuthzMap reduced the time to completion by about 218%, compared to List, and about 165%, compared to Search.

The result of the accuracy analysis (Table 9) rejected two of the null hypothesis for comparing AuthzMap and List, but did not reject any of the hypotheses for comparing AuthzMap and Search.

Table 9: Comparing the correctness of participants' choices in SoD violation detection task.

	A	L	S	A=L	A=S
SoD (R36, R11)	92.5%	83%	91.7%	<0.05	1
SoD (R14, R00)	94.8%	85.9%	89.3%	<0.05	0.451

Application Review Task: Table 4 shows that AuthzMap and List did similarly in terms of efficiency, while Search did worse. In terms of practical significance, AuthzMap reduced the time to completion by about 35%, compared to Search. The accuracy analysis (Table 10) rejected all the null hypotheses for comparing AuthzMap and List in favor of List, and rejected one of the 15 hypotheses for comparing AuthzMap and Search in favor of Search.

Comprehension Task: Our analysis (Table 4) suggests that AuthzMap does better in terms of efficiency than the two other interfaces. It also practically improves efficiency by about 72%, compared to List, and by 89%, compared to Search. This task involved the assessment of risk for users having specific access privileges.

Table 10: Comparing the correctness of participants' choices in application review task.

	A	L	S	A=L	A=S
EdmundJ, R10	82.8%	97%	78.5%	<0.05	1
EdmundJ, R15	81.6%	96.3%	86%	<0.05	1
EdmundJ, R23	81%	97%	76.9%	<0.05	1
EdmundJ, R11	83.3%	97%	80.2%	<0.05	1
EdmundJ, R22	83.3%	97%	78.5%	<0.05	1
EdmundJ, R30	70.7%	97%	86.8%	<0.05	<0.05
EdmundJ, R28	69%	97%	78.5%	<0.05	1
NelsonM, R10	82.2%	97%	78.5%	<0.05	1
NelsonM, R15	82.8%	97%	86%	<0.05	1
NelsonM, R23	79.9%	96.3%	78.5%	<0.05	1
NelsonM, R33	69.5%	96.3%	78.5%	<0.05	1
NelsonM, R35	70.7%	95.6%	85.1%	<0.05	0.149
JaneH, R10	83.3%	96.3%	81.8%	<0.05	1
JaneH, R23	82.8%	97%	81%	<0.05	1
JaneH, R35	71.3%	95.6%	86%	<0.05	0.0909

The summary of participants' responses to risk assessment questions is presented in Figure 7. We used pair-wise two-sided fisher's exact tests with Bonferroni correction, to test the following hypothesis for each of the risk assessment: (A=L) The choice of AuthzMap or List does not impact the accuracy of risk assessment. (A=S) The choice of AuthzMap or Search does not impact the accuracy of risk assessment. The result of the test rejected ($p < 0.05$) the all four (A=L) hypotheses, and rejected ($p < 0.05$) three of the (A=S) hypotheses (in risk assessment of R02, R09, and R11).

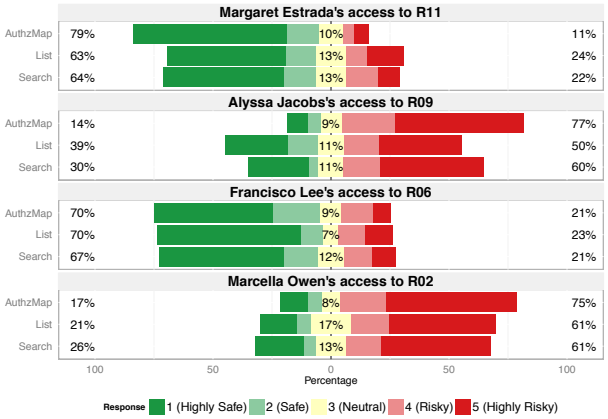


Figure 7: Summary of participants responses to comprehension questions.

Table 4: Median time to completion (TTC) for each of the tasks (in seconds), and pairwise comparison of TTCs. The highlighted cells show the cases where the null hypothesis was rejected and TTC for AuthzMap participants was lower than the other interface.

	Task	A	L	S	A = L	A = S	S=L
1	Training	192.5	243.0	259.0	$p < 0.01, r = 0.24$	$p < 0.01, r = 0.22$	-
2	Common Review	117.5	144.0	96.0	-	$p = 0.01, r = 0.10$	$p < 0.01, r = 0.14$
3	User Comparison	109.5	225.0	195.0	$p < 0.01, r = 0.45$	$p < 0.01, r = 0.37$	-
4	Privilege Accumulation	89.5	256.0	190.0	$p < 0.01, r = 0.50$	$p < 0.01, r = 0.42$	$p < 0.01, r = 0.15$
5	SoD Violation Detection	92.0	293.0	165.0	$p < 0.01, r = 0.57$	$p < 0.01, r = 0.35$	$p < 0.01, r = 0.39$
6	Application Review	181.0	185.0	280.0	-	$p < 0.01, r = 0.34$	$p < 0.01, r = 0.33$
7	Comprehension	247.5	426.0	469.0	$p < 0.01, r = 0.30$	$p < 0.01, r = 0.32$	-

7. DISCUSSION

In this section, we summarize, interpret, and discuss the findings of the user study. We first discuss the efficiency, and accuracy findings. Then, we discuss the limitations of the user study, including the use of non-expert participants, and a synthetic dataset. Finally, we discuss the larger implications of our findings.

7.1 User study findings

7.1.1 Efficiency

In Section 6.3, we show that participants in AuthzMap condition could finish the study faster than those in two other conditions. We also compared the use of three interfaces in various access review scenarios. We showed that AuthzMap improved the efficiency, compared to both of the other interfaces in five of the seven tasks, and compared to one of the interfaces in the two other remaining tasks. This finding require further discussion.

AuthzMap participants' performance in *Common Review* was not as efficient as Search, but was more efficient than List. We can provide three explanations for this: (1) The task involved reviewing only one user. The additional contextual information in the AuthzMap fisheye view could increase user's cognitive load, and hence reduce the performance. (2) The Search interface by default set the status of files to "certify". This helped the participants to change the status of two unauthorized files, and keep the rest of the files intact. Meanwhile, AuthzMap participants had to explicitly set the review status of each file. These two suspected issues provide an opportunity for further improvement. To address the first issue, we can use the focus plus context visualization [23] to highlight the user that the reviewer is currently working on, while still showing the contextual information in the background (e.g., by highlighting the current user and fading the rest of the users). The second possible issue was a design decision that we made in AuthzMap to prevent reviewers from using the default option, and rather make an explicit decision for each access privilege.

The *User Comparison* task was similar to *Common Review*, but it involved three users with identical jobs instead of one. AuthzMap participants did better than participants in two other conditions. We attribute the improvement to AuthzMap's ability to categorize, and filter users according to their job, and then use the contextual information (access of users with the same job) to quickly find the excessive access. Our analysis of study logs confirmed this, as participants used the *sort user* feature of the AuthzMap in this task significantly more than in other tasks. Furthermore, comparing the TTC of this task to the TTC of *Common Review* shows that increasing the number of participants did not impose an additional burden on AuthzMap participants, unlike Search and List participants.

In the *Privilege Accumulation*, *SoD Violation Detection*, and *Comprehension* tasks, AuthzMap performed better than the other two interfaces. We can attribute this to the visibility of context and history in the interface. For example, AuthzMap integrates the employment history and access privileges, and makes it accessible to users. The two other interfaces required participants to collect information from the HR and access review systems, and perform a mental process to formulate the relationship between access control and employment data. Additionally, AuthzMap integrated the SoD policy information with the existing access control data, and helped users quickly identify and resolve the SoD violations. Participants in two other conditions had to use an SoD catalog. Therefore, they needed to mentally associate the policy with the access control data.

In *Application Review* task, AuthzMap participants performed similar to List and better than Search. This is an expected result as both List and AuthzMap clearly integrate information about the application in the interface, but Search participants had to use the auxiliary application catalog to find the files related to a certain application.

7.1.2 Accuracy

In Section 6.3, we report that AuthzMap participants achieved more accurate results than the other two interfaces in only one task.

Accuracy results of *Common Review* task were unexpected. AuthzMap participants committed significantly more errors than participants in two other conditions. Examining the data closely shows many of these participants committed identical errors. After further investigation, we realized that AuthzMap's detail interface showed the user had accumulated privileges from a prior job. And a subset of participants who received the *Privilege Accumulation* task before *Common Review*, did not use the information in File Catalog, but rather did access review based on what they learned from *Privilege Accumulation* (about 15% of the participants in AuthzMap condition). This was our mistake in designing task data, and we should have controlled the privilege accumulation in the policy for this task. If we count the correct answers from the participants who looked at the task from privilege accumulation perspective as valid, there is no statistically significant difference between three conditions in accuracy.

Accuracy results for "User Comparison" task do not show a difference between three conditions. These results suggest the increase in efficiency did not impact the accuracy of participants in AuthzMap condition.

For two tasks that required decision making in uncertain conditions, *Privilege Accumulation* and *SoD Violation Detection*, AuthzMap positively affected accuracy, compared to List but not Search. These two tasks required contextual information that unlike AuthzMap was not integrated in List and Search interfaces. Search participants did surprisingly well in the collection and integration of the context with the information available in the interface but not the Lits participants. One explanation for this observation is that the List interface contains redundant information that could mentally overload users. On the other hand, Search is rather straightforward, and while it requires user to spend more time collecting and integrating information, it does not reduce the accuracy.

Accuracy results for "Application" task were rather surprising. AuthzMap and Search participants produced less accurate results than List. While we expected the List participants to do better than Search (List clearly showed the application associated with each access privileges, as one of the columns in the list of privileges), we expected AuthzMap to perform as good as List. Further looking at the participants errors, we did not find any patterns or evidence that participants committed mistakes rather than slips. There are three possible explanations of such slips: (1) The names of the applications were presented in a small text, and it was rotated 90 degrees. Prior research shows that text rotation can have a negative impact on human cognition, and requires mental rotation, before a human can recognize an object [18]. To address this, we can use slightly less rotated text (e.g., 45 degrees), as it is shown that rotation is positively correlated with cognitive load. (2) Complexity of the grid: to complete the task, users had to recognize the file related to an application (located in columns of the grid), and then check the target user for having access to the file. This process can be prone to errors due to the proximity of grid cells. To address this, we can utilize the focus plus context visualization [23], by allowing users click on the column to focus on a specific file.

Unlike other tasks, AuthzMap participants provided more accurate responses to three of the four questions in the *Comprehension* task. We expected this result, as participants in other conditions

should have used multiple information sources to complete the task, and they needed to build the correct model of the policy in their memory. Yet, AuthzMap participants could see the complete picture of the policy. The only question that we did not see a significant difference between AuthzMap and both conditions was the assessment of R06 risk, which we did not see a difference between AuthzMap and Search. R06 could be both safe or unsafe, therefore, we expect participants not to choose highly safe or highly risky. Further look at the graphs in Figure 7 shows that Search participants assessed R06 and R11 (which was highly safe) similarly. However, AuthzMap participants assessed R06 rather differently from how they assessed R11. This suggests that maybe Search participants naively chose unsure responses, but AuthzMap participants made a more informed choice.

7.1.3 User Study Limitations

Ideally we would have evaluated AuthzMap by asking managers to use AuthzMap to review access of actual users in their company. But our experience from this study, and our past field studies [3] suggests that conducting a field experiments in real organizations is extremely difficult. In this study, we faced challenges similar to those discussed in [31]. First, AuthzMap is a prototype, and integrating it with real access management systems in organizations is a software engineering challenge. Second, asking managers to budget time for evaluating AuthzMap is challenging, particularly because access review is not their day-to-day task. Third, AuthzMap requires identity and access control data, which are commonly considered extremely sensitive. Our experience shows that even getting permission to conduct an interview requires approval from the legal department of a large company, as well as multiple managers, let alone conducting experiments using the sensitive data.

Due to the above challenges, we adopted an approach similar to [31], and conducted a set of during-design, exploratory studies before committing to a costly field study. First, we received feedback on AuthzMap from a large domain expert audience (employees of our industry partner). We also had two small group discussions with the engineering team, and usability team of our industry partner. Second, we conducted 12 heuristic evaluation sessions (using Nielsen [25] and ITSM [17] heuristics) with independent usability experts to identify usability issues with AuthzMap, and further improved the interface. Third, we conducted a lab study (Section 6) with non-domain experts to further evaluate the interface and compare it to existing systems. Sedlmair et al. [31] showed that conducting during-design experiments could be very helpful and lead to tools with higher usability, and eventually become a major reason for the tool being deployed in the field. Therefore, we conducted an exploratory study with MTurk participants to be confident that the tool does not have obvious usability problems, and fares well against existing systems.

The next step in evaluation would be to conduct an in-depth long-term case study [32] in an organization, by integrating AuthzMap with existing access management systems, asking managers to use AuthzMap, and then get qualitative feedback on the impact of AuthzMap. Such a field study can show if the tool will be adopted by managers, and could increase the effectiveness of conducting access reviews.

We used an automatically generated dataset. Using a real-world dataset was not feasible, as there are very few real-world enterprise access control data sets available to the research community. We examined five common datasets used regularly by access control community such as: americas_small, apj, healthcare, domino, firewall1 and firewall2 [11]. These datasets only contained lists of users, permissions, and user-to-permission assignments. Our study required contextual data, such as users' job, employment history, access history, and review history. Adding meaningful context to existing datasets was not possible, therefore, we elect to generate a dataset that best matched our interview study findings.

7.2 Implications Beyond Access Review

Our field study findings have larger implications than just understanding access review activity. Our findings suggests that while access control policies are usually composed of users, roles, and permissions, these three components are only parts of a larger context, and they evolve and change over time. Therefore, a snapshot of a user's access privileges does not provide a complete picture of access policy. We further determined the context of a users' access privileges, which includes other users' access, other policies that impact such access (such as SoD policies), user's job, and other stakeholders involved in the access control decisions, such as those who requested or approved the user's access. We also demonstrated that access control policies evolve over time, and identified users' job, access privileges, and previous reviews as important historical artifacts. Although our focus was on access control in large organizations, the concept of context for access control policies is still applicable to access control in other domains such as file systems, multimedia, etc. We should note that each domain should be studied separately, as the contextual information for enterprise domain (such as job or approval workflow) may not be applicable in other domains. For example, findings by Vaniea et al. [36] suggest that proximity of access control displays and photos helps users notice and correct access control errors. In this case, the photo (visual representation of the asset) is a part of the access control context.

The design of AuthzMap can serve as an example of how the contextual information can be integrated with access policy in a user interface, and our user study suggested that the design was successful. Furthermore, such integration will improve efficiency of accessing contextual information, and in complex decision making processes (such as *Comprehension* task in our study) can improve better understanding of policy, and therefore, facilitate making more accurate decisions. Our study results also suggest that showing context could increase the complexity of the interface and in few occasions could negatively impact the accuracy or efficiency. Therefore, we suggest improvements such as focus plus context visualization [23] to alleviate those conditions.

8. CONCLUSION

In this paper, we studied how access policies are reviewed in large organizations. We then identified a set of five challenges that organizations face during access review, and suggested four design goals to deal with those challenges. We then realized the design goals by building AuthzMap, a novel user interface for reviwing and making sense of access policies in organizations. We then conducted an exploratory user study with 340 MTurk participants to compare the use of AuthzMap to two of the existing access review systems. Our results show that AuthzMap improved efficiency of access review in five of the seven, and accuracy in one of the seven tasks. Our goal for designing AuthzMap was to address five challenges identified during the field study, and our results show that for those tasks that involve identified challenges, AuthzMap improved the efficiency, and in one task accuracy. The bigger HCI implications of this work are exploring the importance of context in access control, and proposing an effective approach for integrating contextual information in access control interfaces. As the next step, AuthzMap should be deployed in a real organizational setting, and its impact should be evaluated in a field study.

9. REFERENCES

[1] L. Bauer, L. F. Cranor, R. W. Reeder, M. K. Reiter, and K. Vaniea. Real life challenges in access-control management. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 899–908, New York, NY, USA, 2009. ACM.

[2] M. Beckerle and L. A. Martucci. Formal definitions for usable access control rule sets from goals to metrics. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 2:1–2:11, New York, NY, USA, 2013. ACM.

[3] D. Botta, R. Werlinger, A. Gagné, K. Beznosov, L. Iverson, S. Fels, and B. Fisher. Towards understanding IT security professionals and their tools. In *Proc. of Symp. On Usable Privacy and Security (SOUPS)*, pages 100–111, Pittsburgh, PA, July 18-20 2007.

[4] C. Brodie, C.-M. Karat, J. Karat, and J. Feng. Usable security and privacy: a case study of developing privacy management tools. In *SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security*, pages 35–43, New York, NY, USA, 2005. ACM.

[5] J. M. Carroll, P. L. Smith-Kerker, J. R. Ford, and S. A. Mazur-Rimetz. The minimal manual. *Human–Computer Interaction*, 3(2):123–153, 1987.

[6] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/, 1996.

[7] K. Charmaz. *Constructing Grounded Theory*. SAGE publications, 2006.

[8] J. Considine, M. Botti, and S. Thomas. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12(1):19 – 24, 2005.

[9] G. Convertino, H. M. Mentis, A. Slavkovic, M. B. Rosson, and J. M. Carroll. Supporting common ground and awareness in emergency management planning: A design research project. *ACM Trans. Comput.-Hum. Interact.*, 18(4):22:1–22:34, Dec. 2011.

[10] A. Cser. The forrester wave: Role management and access recertification, q3 2011. Technical report, Forrester Research, inc., August 2011.

[11] A. Ene, W. Horne, N. Milosavljevic, P. Rao, R. Schreiber, and R. E. Tarjan. Fast exact and heuristic methods for role minimization problems. In *SACMAT '08: Proceedings of the 13th ACM symposium on Access control models and technologies*, pages 1–10, New York, NY, USA, 2008. ACM.

[12] Y. Engeström. Activity theory and individual and social transformation. *Perspectives on activity theory*, pages 19–38, 1999.

[13] A. Forget, S. Chiasson, and R. Biddle. Supporting learning of an unfamiliar authentication scheme. In *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, volume 2012, pages 1002–1011, 2012.

[14] J. Hollan, E. Hutchins, and D. Kirsh. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.*, 7(2):174–196, 2000.

[15] P. Inglesant, M. A. Sasse, D. Chadwick, and L. L. Shi. Expressions of expertness: the virtuous circle of natural language for access control policy specification. In *SOUPS '08: Proceedings of the 4th symposium on Usable privacy and security*, pages 77–88, New York, NY, USA, 2008. ACM.

[16] P. Jaferian and K. Beznosov. Access review survey report. Technical Report LERSSE-REPORT-2014-001, Laboratory for Education and Research in Secure Systems Engineering, University of British Columbia, May 2014.

[17] P. Jaferian, K. Hawkey, A. Sotirakopoulos, M. Velez-Rojas, and K. Beznosov. Heuristics for evaluating it security management tools. *Human–Computer Interaction*, 29(4):1–40, 2013.

[18] P. Jolicoeur. The time to name disoriented natural objects. *Memory & Cognition*, 13(4):289–303, 1985.

[19] V. Kaptelinin and B. Nardi. *Acting with technology: Activity theory and interaction design*. MIT Press, 2006.

[20] V. Kaptelinin, B. A. Nardi, and C. Macaulay. Methods & tools: The activity checklist: a tool for representing the space of context. *interactions*, 6(4):27–39, July 1999.

[21] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[22] B. W. Lampson. Protection. In *5th Princeton Conference on Information Sciences and Systems*, page 437, New York, NY, USA, 1971. ACM Press.

[23] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Trans. Comput.-Hum. Interact.*, 1(2):126–160, June 1994.

[24] M. McCloskey. Turn user goals into task scenarios for usability testing, January 2014.

[25] J. Nielsen. Finding usability problems through heuristic evaluation. In *Proc. CHI '92*, pages 373–380, New York, NY, USA, 1992. ACM.

[26] J. Nielsen. Usability 101: Introduction to usability. http://www.nngroup.com/articles/usability-101-introduction-to-usability/, January 2012.

[27] D. A. Norman. *The Psychology of Everyday Things*. Basic Books, 1988.

[28] D. Povey. Optimistic security: A new access control paradigm. In *Proceedings of the 1999 Workshop on New Security Paradigms*, NSPW '99, pages 40–45, New York, NY, USA, 2000. ACM.

[29] R. W. Reeder, L. Bauer, L. F. Cranor, M. K. Reiter, K. Bacon, K. How, and H. Strong. Expandable grids for visualizing and authoring computer security policies. In *Proc. CHI '08*, pages 1473–1482, New York, NY, USA, 2008. ACM.

[30] R. W. Reeder, L. Bauer, L. F. Cranor, M. K. Reiter, and K. Vaniea. More than skin deep: measuring effects of the underlying model on access-control system usability. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 2065–2074, New York, NY, USA, 2011. ACM.

[31] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz. Information visualization evaluation in large companies: Challenges, experiences and recommendations. *Information Visualization*, 10(3):248–266, July 2011.

[32] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, pages 1–7, New York, NY, USA, 2006. ACM.

[33] D. K. Smetters and N. Good. How users use access control. In *SOUPS '09: Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, New York, NY, USA, 2009. ACM.

[34] T. Swensen. Wikileaked! wikileaked! what we can all learn from the bradley manning debacle. Novell connections magazine, April 2011.

[35] Unknown. Sarbanes-Oxley Act of 2002. Online Document, July 2002.

[36] K. Vaniea, L. Bauer, L. F. Cranor, and M. K. Reiter. Out of sight, out of mind: Effects of displaying access-control information near the item it controls. In *Proceedings of the 2012 Tenth Annual International Conference on Privacy, Security and Trust (PST)*, PST '12, pages 128–136, Washington, DC, USA, 2012. IEEE Computer Society.

[37] K. Vaniea, L. Bauer, L. F. Cranor, and M. K. Reiter. Studying access-control usability in the lab: lessons learned from four studies. In *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results*, LASER '12, pages 31–40, New York, NY, USA, 2012. ACM.

[38] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. 'i regretted the minute i pressed share': a qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 10:1–10:16, New York, NY, USA, 2011. ACM.

APPENDIX

A. INTERVIEW GUIDE

A.1 Organizational context

A.1.1 General Information about Interviewee and Organization

- What is your position?
- Background: What is your IT/Security education/path?
- Can you briefly describe your organization? (size, sector)
- Describe security management within your organization
  - Who is responsible for security within your organization?
  - What is the security management model (centralized, distributed, etc.)? (With little help to the person)
- Can you describe the security policies in your organization (also probe for participant's role)
  - What formal (official, written) security guidelines/ policies/ architectures/ models are in place?
  - What is done in practice? (To see if the policy is completely enforced)
  - What is the process for developing policies?
  - How are policies communicated?
- To whom are policies communicated?
  - How are security-related policies enforced?
- What security risks/challenges do you perceive to be important for your organization?
  - What are the security risks or challenges in your organization?
  - What security incidents has your organization experienced as a result of these risks/challenges?
  - To what extent these incidents relate to access and identity management?
  - Are there security incidents or risks that are least priority?

A.1.2 Activities

- What are your responsibilities within the organization? (get overall, lead into security specific activities)
  - Actual duties/ official duties (Let them talk, probe anything not on list to confirm that omissions are true negatives)
- Manage identities and accesses
- Perform and respond to security audits on the IT infrastructure?
- Develop security policies?
- Design and revise security services or projects?
- Implement security controls?
- Solve end user security issues?
- Educate and train?
- Respond to security incidents? (Skills, knowledge and strategies, resources (tools) used)
- Mitigate new security vulnerabilities?
- Prioritization (typical day)

A.2 Questions about Access and Identify Management (AIM) Process

A.2.1 AIM process (general)

- What do you consider to fall under the definition of access and identity management?
- What is the current process within your organization?
  - Activities? (policies, managing access, managing identities, audit, compliance, trouble shooting)
- Stakeholders? (management, HR, IT, security, employees, customers, external organizations...)
- What is your role?
- Knowledge required
- Importance?
- Frequency?
- Is it supported by tools?
- Can it be automated or supported better by the tool?
- How was this process before adopting an IdM solution ?

A.2.2 Compliance

- Is the organization required to comply with any standard? Which standard?
- What is the role of IDM solution in your compliance with the standard?

A.3 Probing specific activities (depends on their role)

A.3.1 Managing accesses and identities

- Can you describe the lifecycle for managing accesses and identities? (From creation to destruction of an identity)
- Which parts of this lifecycle is supported by your IdM system?
- How you manage changes in user status? (extending access for a user, changing access, discontinuing access)
- How frequently you face exceptions in setting up accesses and how you handle them? (For example: Employees should normally access X but not Y. But for a specific case you should temporarily provide access to an employee to Y.)
- How complex are the policies and how do you handle complexity?
  - Number of users? Number of resources? Number of roles? Number of access rules (E.g. Role X has access Y to resource Z)
- Are there any cases that you don't want system access to be controlled by your IDM solution?

A.3.2 Entitlements

- Can you give us a definition for entitlement ? Can you give us examples from your organization?
- How entitlements are managed in your organization ? Is there a process in place?
- What stakeholders are involved in determining the meaning of an entitlement and deciding about associating entitlements to users ?
- What is the process of checking if users are assigned to a correct set of entitlements?

A.3.3 Audit

- How can you make sure that the correct access rights are set for the intended person? (that the policy is implemented correctly)
- What is the process for identifying and removing the unused or discontinued identities and accesses?
- Do you have any formal audit procedure in place? If so, describe?
- Is there any legislation that require your organization to perform audit ?

A.3.4 Role Management

- How do you create roles in your organization? (define business responsibilities as roles and association of roles to entitlements?
- How frequently roles are changed or added?
- How do you perform "role engineering" in your system?
  - What is difficult/easy about it?
  - What approach do you use (top down, bottom up, hybrid)?
- What stakeholders are involved in the process of managing roles?
- What tools do you use for managing roles?

A.3.5 End-user experience

- What are the ways of accessing the system for users? Is there just one, or many (different usernames, different portals, etc)?
- Can you recall any end-user complaints relating to the IdM solution?
- Is it possible for users to manage access?
- How do the end users understand the configuration implemented by security practitioners? How can an end-user know which resources he has access to?
  - Does the tool give feedback?
  - Do you need to provide explicit knowledge? (For example about how they can find-out their access rights, changing their personal information (password, etc.))?
  - Do end-users need to be aware of their access rights or policy at all?
- Do you think the end-user experience has changed after adoption of IdM system ?

A.3.6 Troubleshooting

- How frequently you deal with problems that require troubleshooting?
- Can you give an example? (get details: collaboration?, blow-by-blow account)
- While performing troubleshooting, what is the magnitude of information that you work with? (means logs about accesses) Do you cut things or prioritize because of the volume of information?

A.3.7 Archiving

- What kind of activities/incidents/interactions/communications do you document and how?
- Is there a need for recording/archiving of communications? In what circumstances?

A.3.8 Reporting

- Describe the reports that you generate that are related to access and identity management.
- For whom do you generate these reports?
- How are your reports used?
- What tools do you use to help compose and send your IdM reports?
- Do you generate reports for different people? Who?
- If you compose different kinds of reports (different content, different level of granularity) for different people, is it easy for you to compose different kinds?
- What makes it easy or tedious?
- Do any of your report help you prioritize? What information helps? Where does it come from?

A.4 Questions about Access and Identify Management (AIM) Technologies

- What is your definition of an ideal IdM solution? (Solution that manage accesses, control digital identities, enable checking who did what and who granted the access, checking the compliance of the system)
- Do you currently have such solution?
- Which parts exist in your current infrastructure?
- What are the driving forces for adopting IdM technology in your organization ?

A.4.1 Purchasing/Evaluation

- What was the process for selecting the IdM tool in use?
  - What stakeholders are involved in the process?
  - How did you evaluate the competing tools?
- What features do you look for in a tool? Which features are available in your current tools?
- What properties to you wish for in your tools? (quality, user interface, performance, service, vendor reputation).

A.4.2 Tool deployment

- What are the pre-requisites for deploying an IdM solution? I mean should any specific business processes in place? Should any technological infrastructure be in place? Is there any training required? Is there any kind of knowledge required?
- Who are the people involved in the IdM deployment? I mean is there any relation for example with managers, end-users, or external organizations?
- What are the difficulties in deployment of the product?
- Do you need to customize out of the box identity and access management tools to meet your needs? If yes, can you describe the process for that?
- Do you need to integrate any of your existing systems (Databases, Terminals, Web Applications, etc.) with your IdM solution? Does the solution perform this automatically?
- Do you have any recommendations for improving deployment process?

A.4.3 Tool maintenance

- What maintenance tasks do you perform to keep the IdM solution running and who is responsible for them?
- How much technical knowledge and effort do they need to maintain the solution?
- What is the process of updating or changing your IdM solution?

A.4.4 Tool Use

- How do you use tool X and what do you like/dislike about it? (if possible, get them to show the interface and probe their view of the functionality/usability afforded by the tool. Try to take photos or draw sketches from what the show.)
- In addition to tools that are part of your general IdM infrastructure, are there any other tools used for the various IdM activities? (i.e., excel sheet for creating reports related to IdM)
- Are there any tools do you no longer use? (why?)
- What is the most error prone part of your identity management solution?
  - How do you find out that a tool has made an error?
  - What do you do to recover from errors?

A.5 Working/Dealing with other stakeholders

A.5.1 Collaboration

- With whom do you interact during IdM activities? What are the circumstances?
- Do you need to Co-ordinate your work with other people?
  - Do you need to delegate some part of an IdM task to other people? Do you need to work with other people in order to accomplish an IdM task?
- What is your relationship with other people who are responsible for identity management? How closely do you work with them?
- Do the people who manage accesses or identities have knowledge about computer security? Do they know whether or not risks are involved in what they do? Do they understand these risks?
  - Tools to facilitate awareness: Do you use any tools to support awareness of activities of others (workflows, shared calendars, shared to-do lists, whiteboards)
  - Does the IdM tool provide any support for activities which require collaboration?

A.5.2 Communication and Common ground (negotiating a shared understanding?

- What type of information do you need to share?
- Are there new issues that arise through your on-going experience with IdM which are necessary to communicate to others?
  - How are they communicated? (Can give example of Documents, Wikis, or SharePoint)
  - Is your IdM tool integrated with any of these communication channels?
  - Do you use specific terminology to communicate with other people involved in IdM activities?
- How do they know that the information and your communication is understood?
- How people understand each other while communicating and how they make sure and let each other know that they understood each other?
- Can you give us an example of misunderstanding during communication with other stakeholders about IdM?
- When is it necessary to interact with people outside of the organization?

B. DETAILED DESCRIPTION OF AUTHZMAP, LIST, AND SEARCH

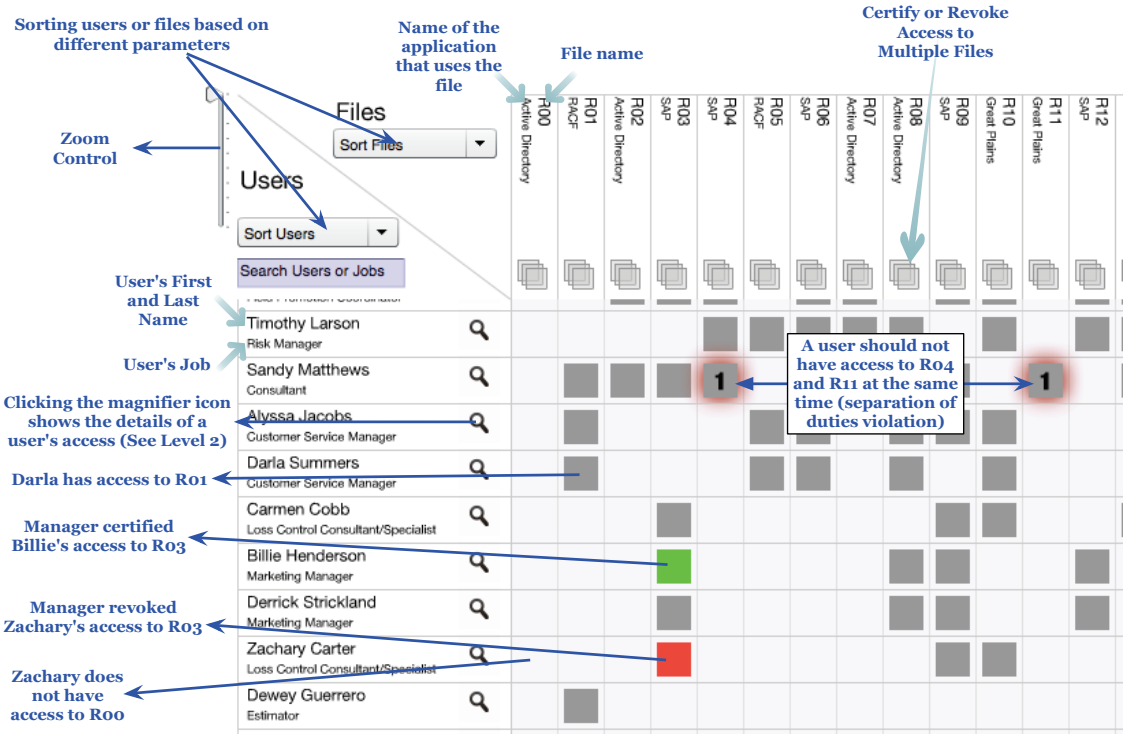


Figure 8: Level one of the AuthzMap interface. We used the notion of files in the user study, but eventually columns in the grid indicate roles, permissions, files, or any other type of entitlements.

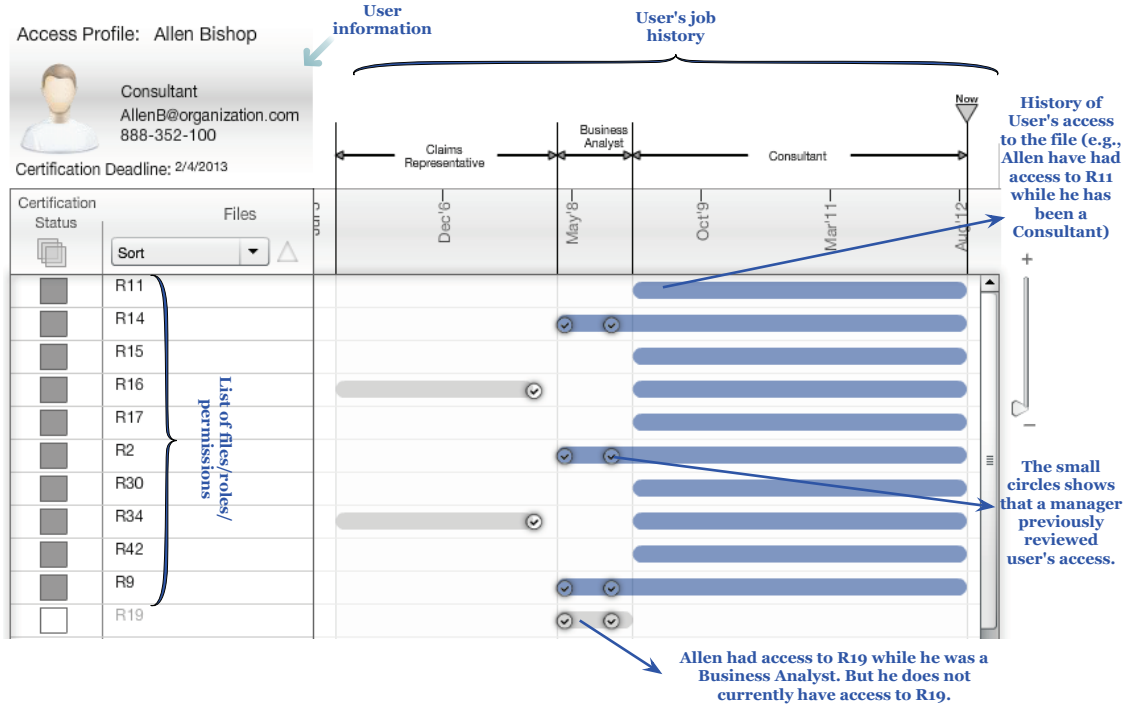


Figure 9: Level two of the AuthzMap interface. Reviewer can access this level by clicking on the magnifier icon in the level 1 of the interface.

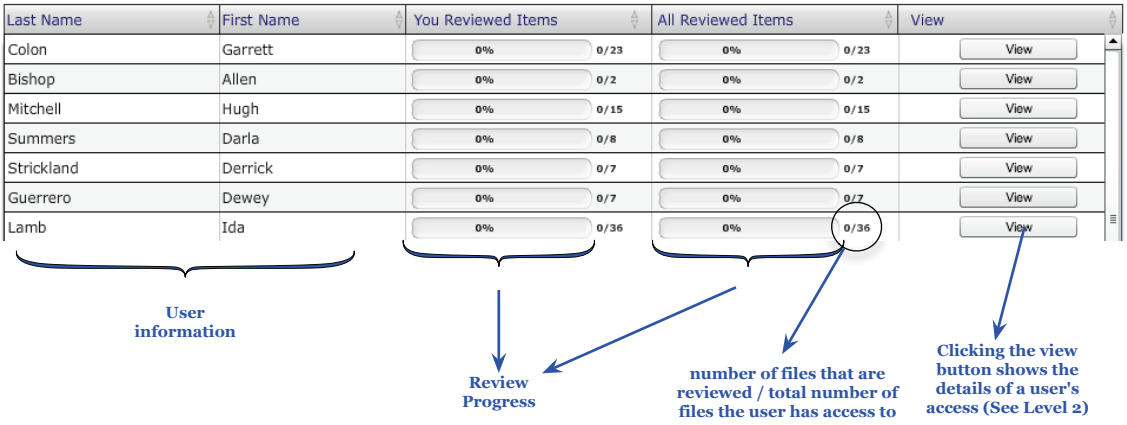


Figure 10: Level one of the List interface. The original interface used the notion of "entitlements", but we changed it to files for the purpose of the user study.

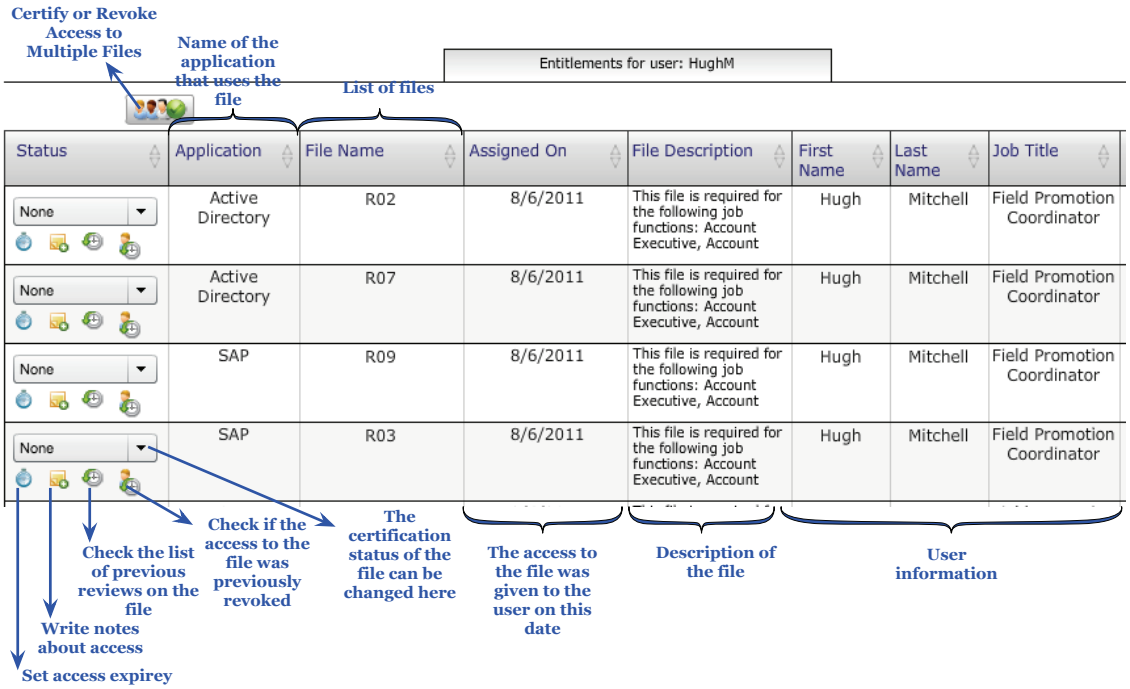


Figure 11: Level two of the List interface.

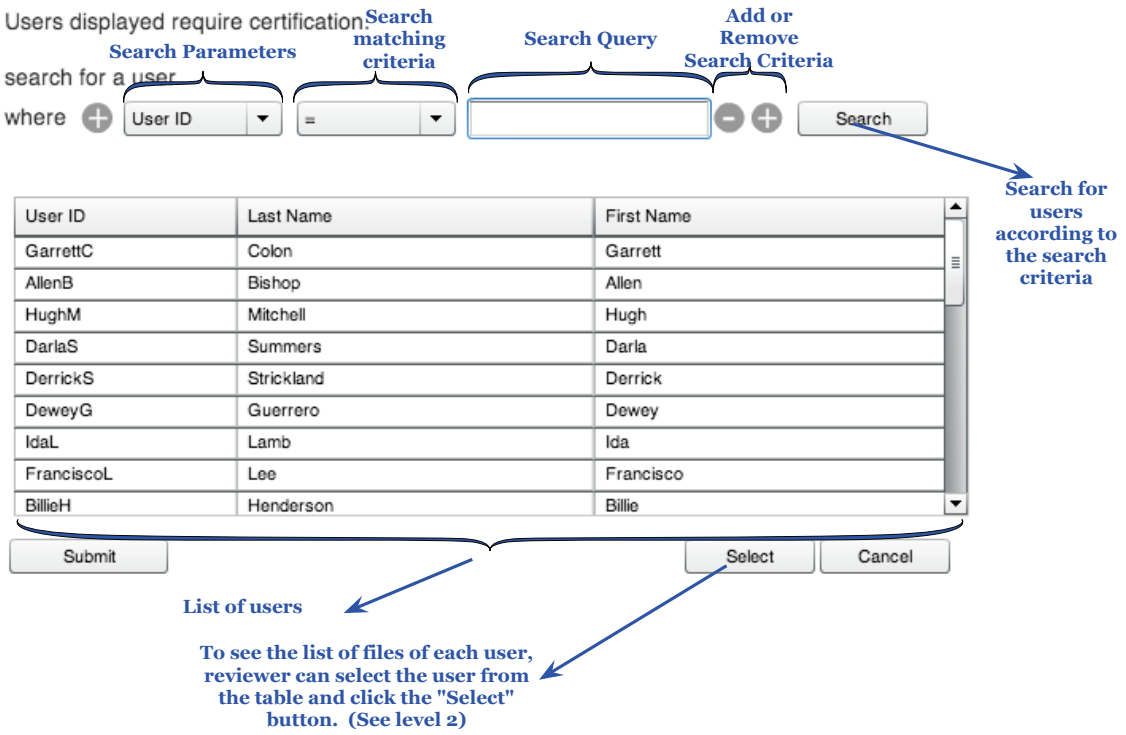


Figure 12: Level one of the Search interface. The original interface used the notion of "Roles", but we changed it to files for the purpose of the user study.

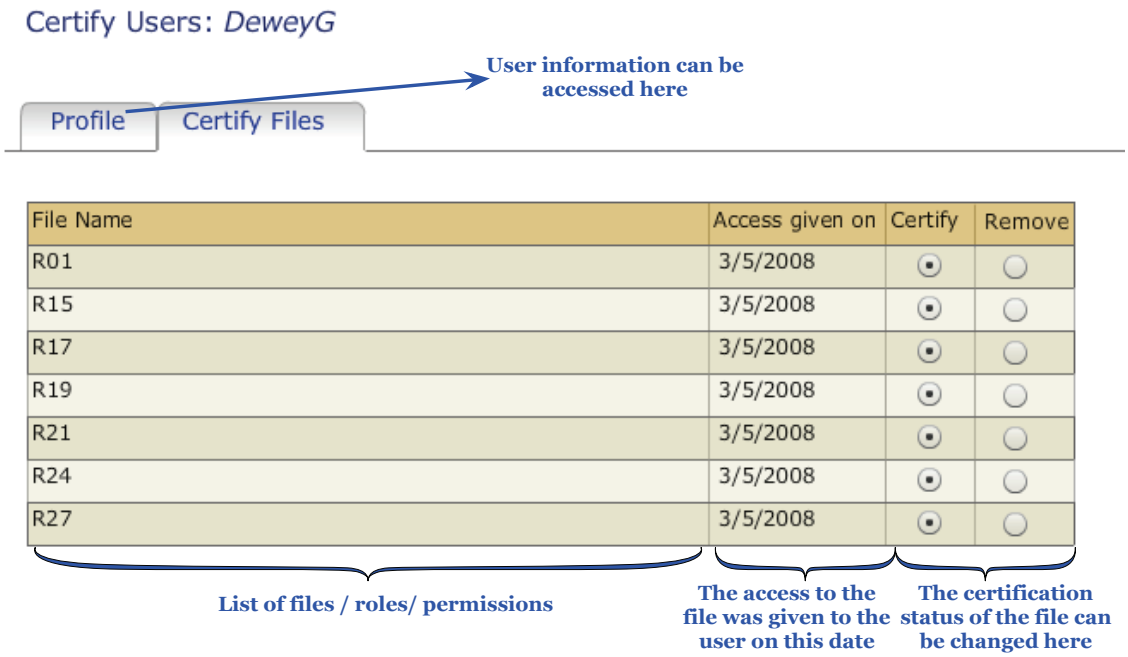


Figure 13: Level two of the Search interface.