Dynamics of Real-world Networks

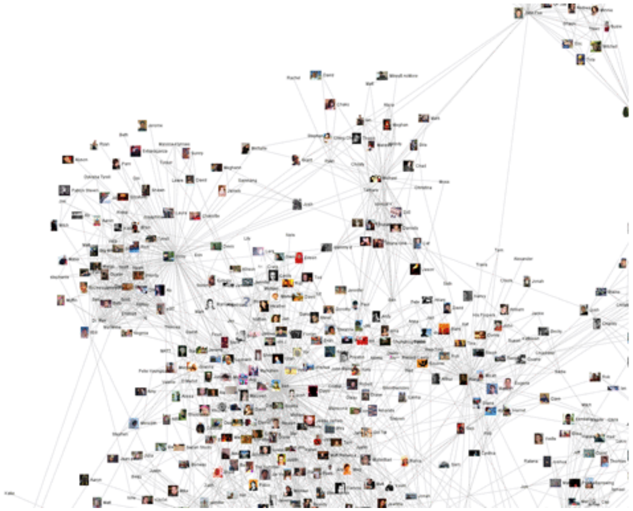
Jure Leskovec
Machine Learning Department
Carnegie Mellon University
jure@cs.cmu.edu
http://www.cs.cmu.edu/~jure

Committee members

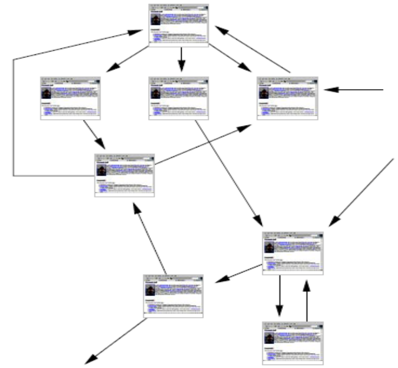
- Christos Faloutsos
- Avrim Blum
- Jon Kleinberg
- John Lafferty



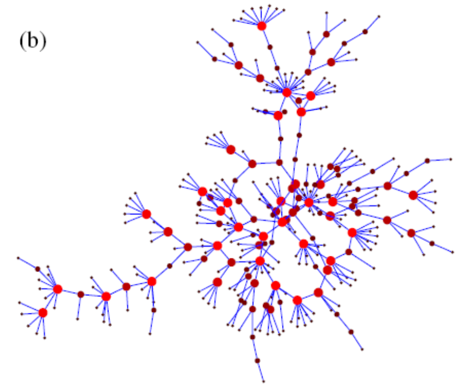

Network dynamics



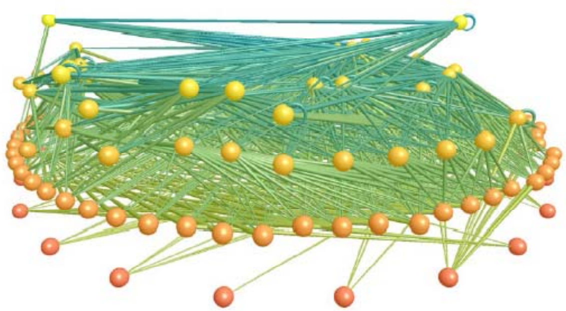
Friendship network



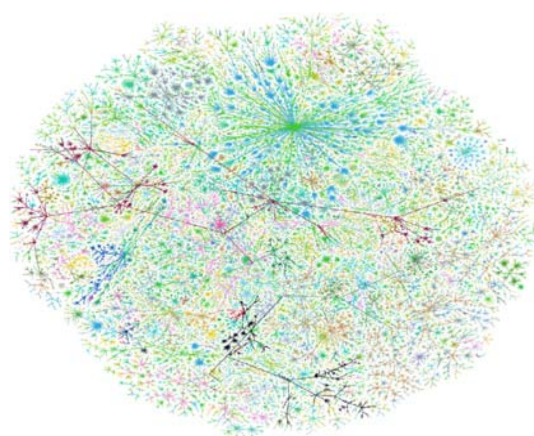
Web & citations



Sexual network



Food-web
(who-eats-whom)



Internet

Yeast protein
interactions

Large real world networks

- Instant messenger network
  - N = 180 million nodes
  - E = 1.3 billion edges
- Blog network
  - N = 2.5 million nodes
  - E = 5 million edges
- Autonomous systems
  - N = 6,500 nodes
  - E = 26,500 edges
- Citation network of physics papers
  - N = 31,000 nodes
  - E = 350,000 edges
- Recommendation network
  - N = 3 million nodes
  - E = 16 million edges

Questions we ask

- Do networks follow patterns as they grow?
- How to generate realistic graphs?
- How does influence spread over the network (chains, stars)?
- How to find/select nodes to detect cascades?

Our work: Network dynamics

- Our research focuses on analyzing and modeling the structure, evolution and dynamics of large real-world networks
  - Evolution
    - Growth and evolution of networks
  - Cascades
    - Processes taking place on networks

Our work: Goals

- 3 parts / goals
  - G1: What are interesting statistical properties of network structure?
    - e.g., 6-degrees
  - G2: What is a good tractable model?
    - e.g., preferential attachment
  - G3: Use models and findings to predict future behavior
    - e.g., node immunization

Our work: Overview

	S1: Dynamics of network evolution	S2: Dynamics of processes on networks
G1: Patterns		
G2: Models		
G3: Predictions		

Our work: Overview

	S1: Dynamics of network evolution	S2: Dynamics of processes on networks
G1: Patterns	KDD ‘05 TKDD ’07	PKDD ‘06 ACM EC ’06
G2: Models	KDD ‘05 PAKDD ’05	SDM ‘07 TWEB ’07
G3: Predictions	KDD ‘06 ICML ’07	WWW ‘07 submission to KDD

Our work: Impact and applications

- Structural properties
  - Abnormality detection
- Graph models
  - Graph generation
  - Graph sampling and extrapolations
  - Anonymization
- Cascades
  - Node selection and targeting
  - Outbreak detection

Outline

- Introduction
- Completed work
  - S1: Network structure and evolution
  - S2: Network cascades
- Proposed work
  - Kronecker time evolving graphs
  - Large online communication networks
  - Links and information cascades
- Conclusion

Completed work: Overview

	S1: Dynamics of network evolution	S2: Dynamics of processes on networks
G1: Patterns	Densification Shrinking diameters	Cascade shape and size
G2: Models	Forest Fire Kronecker graphs	Cascade generation model
G3: Predictions	Estimating Kronecker parameters	Selecting nodes for detecting cascades

Completed work: Overview

	S1: Dynamics of network evolution	S2: Dynamics of processes on networks
G1: Patterns	Densification Shrinking diameters	Cascade shape and size
G2: Models	Forest Fire Kronecker graphs	Cascade generation model
G3: Predictions	Estimating Kronecker parameters	Selecting nodes for detecting cascades

G1 - Patterns: Densification

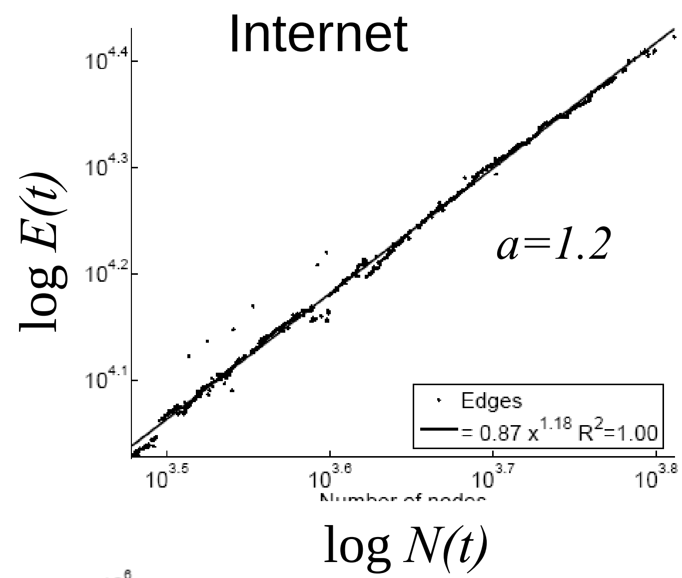
- What is the relation between the number of nodes and the edges over time?
- Networks are denser over time
- Densification Power Law

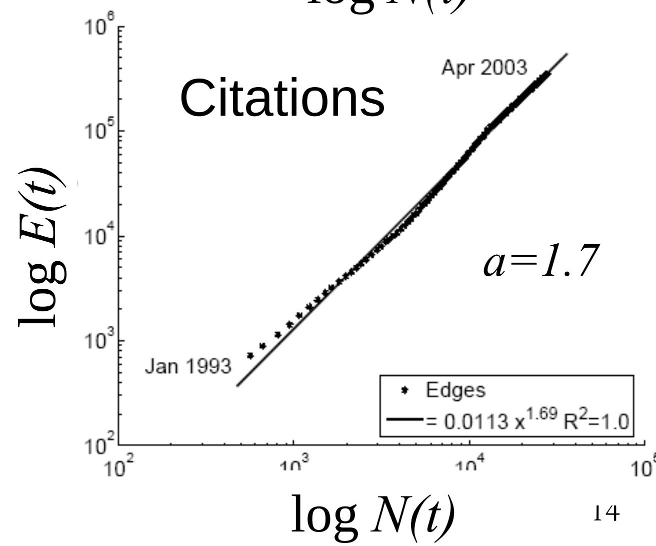
$$E(t) \propto N(t)^a$$

a … densification exponent:

$1 \leq a \leq 2$:

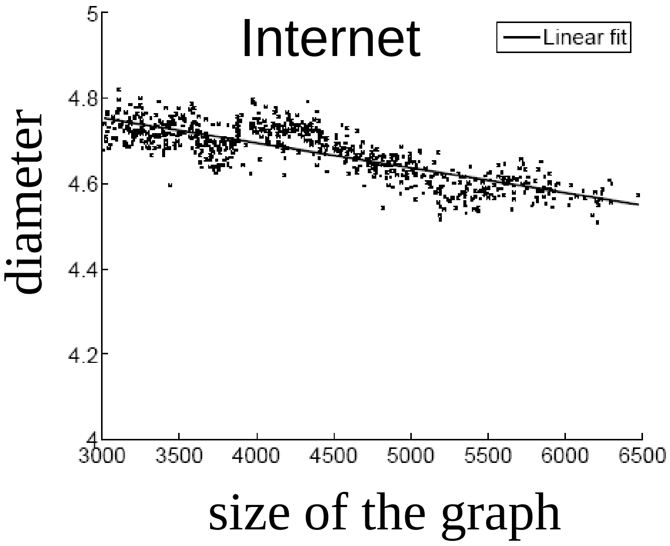
- $a=1$: linear growth – constant degree
- $a=2$: quadratic growth – clique



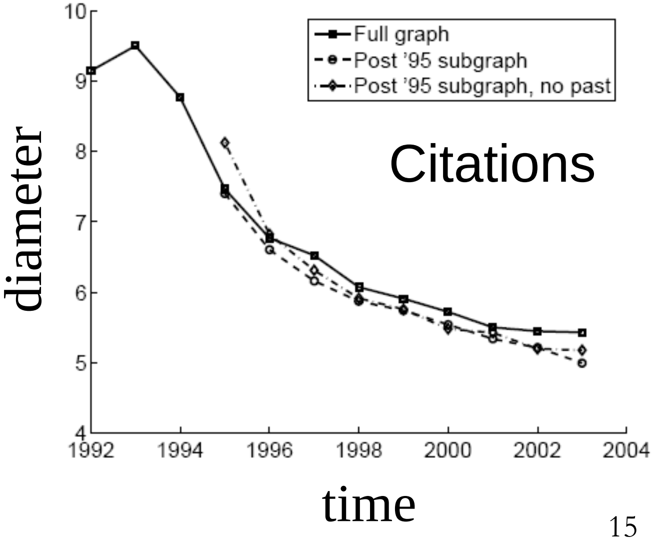


G1 - Patterns: Shrinking diameters

- ~~Intuition and prior work say that distances between the nodes slowly grow as the network grows (like $\log N$)~~



- Diameter Shrinks or Stabilizes over time
  - as the network grows the distances between nodes slowly decrease

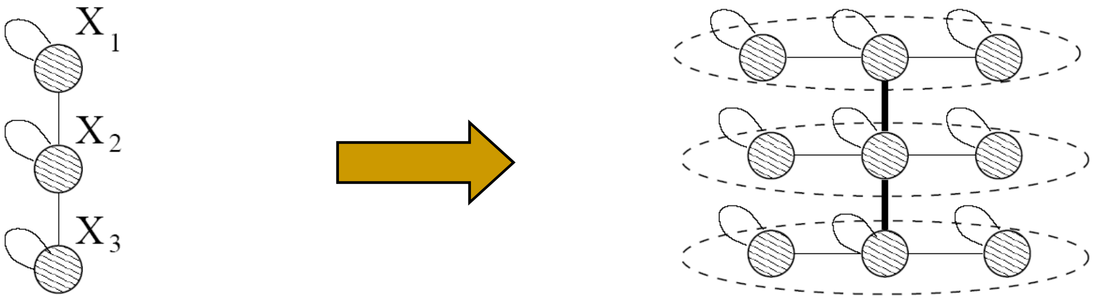


G2 - Models: Kronecker graphs

- Want to have a model that can generate a realistic graph with realistic growth
  - Patterns for static networks
  - Patterns for evolving networks
- The model should be
  - analytically tractable
    - We can prove properties of graphs the model generates
  - computationally tractable
    - We can estimate parameters

Idea: Recursive graph generation

- Try to mimic recursive graph/community growth because self-similarity leads to power-laws
- There are many obvious (but wrong) ways:

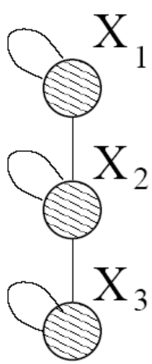


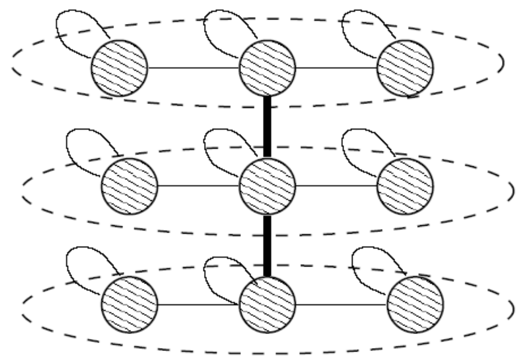
Initial graph

Recursive expansion

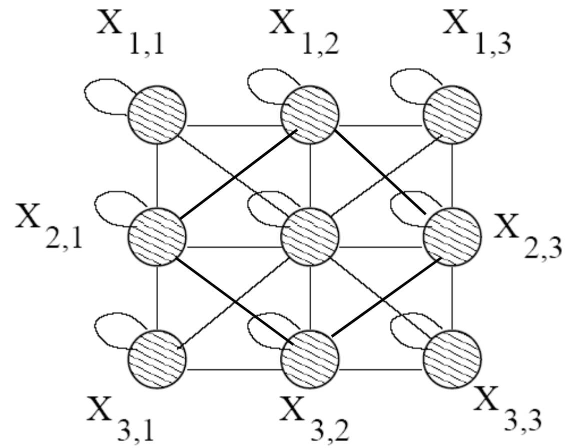
  - Does not densify, has increasing diameter
- Kronecker Product is a way of generating self-similar matrices

Kronecker product: Graph





Intermediate stage



1	1	0
1	1	1
0	1	1

(3x3)

G_1

Adjacency matrix

G_1	G_1	0
G_1	G_1	G_1
0	G_1	G_1

(9x9)

$$G_2 = G_1 \otimes G_1$$

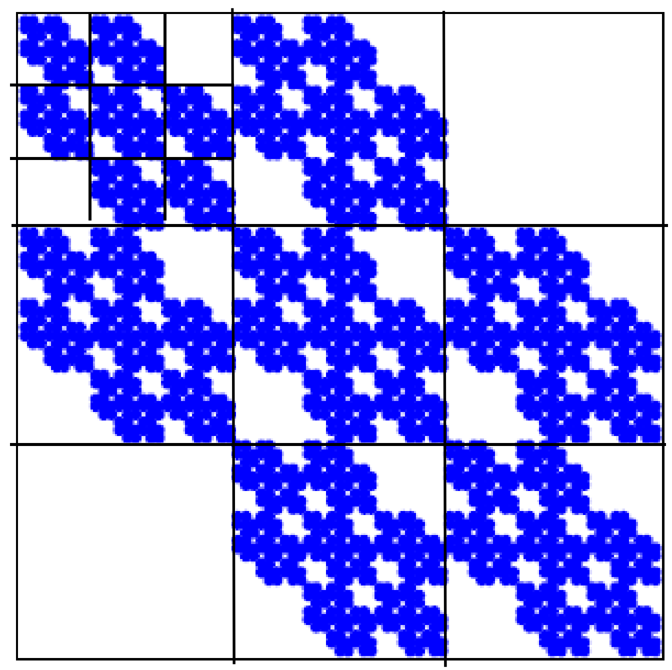
Adjacency matrix

Kronecker product: Graph

- Continuing multiplying with G_1 we obtain G_4 and so on …

1	1	0
1	1	1
0	1	1

G_1



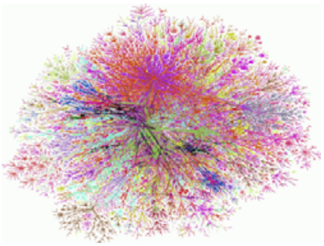
G_4 adjacency matrix

Properties of Kronecker graphs

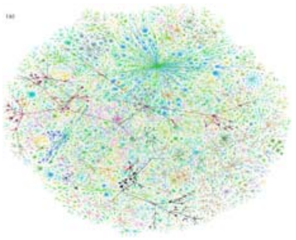
- We show that Kronecker multiplication generates graphs that have:
  - Properties of static networks
    - ✓Power Law Degree Distribution
    - ✓Power Law eigenvalue and eigenvector distribution
    - ✓Small Diameter
  - Properties of dynamic networks
    - ✓ Densification Power Law
    - ✓ Shrinking / Stabilizing Diameter
- This means "shapes" of the distributions match but the properties are not independent
- How do we set the initiator to match the real graph?

G3 - Predictions: The problem

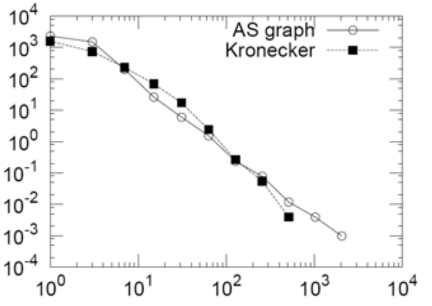
- We want to generate realistic networks:



Given a
real network



Generate a
synthetic network

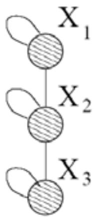


Compare some property,
e.g., degree distribution

  - G1) What are the relevant properties? ✓
  - G2) What is a good tractable model? ✓
  - G3) How can we fit the model (find parameters)?

Model estimation: approach

- Maximum likelihood estimation
  - Given real graph G
  - Estimate the Kronecker initiator graph Θ (e.g., $\begin{array}{|c|c|c|}\hline 1&1&0\\\hline 1&1&1\\\hline 0&1&1\\\hline\end{array}$) which



$$\arg\max_{\Theta} P(G \mid \Theta)$$

- We need to (efficiently) calculate

$$P(G \mid \Theta)$$

- And maximize over Θ

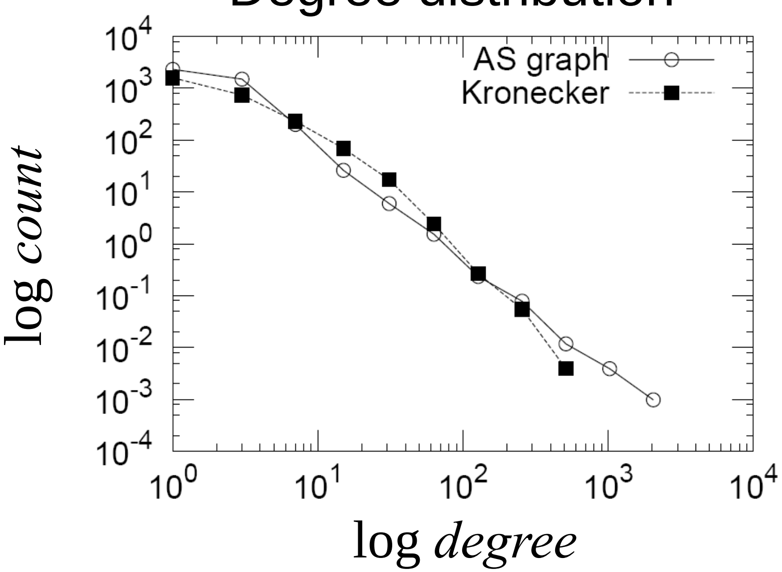
Model estimation: solution

- Naïvely estimating the Kronecker initiator takes $O(N!N^2)$ time:
  - $N!$ for graph isomorphism
    - Metropolis sampling: $N! \rightarrow$ *(big) const*
  - N^2 for traversing the graph adjacency matrix
    - Properties of Kronecker product and sparsity ($E << N^2$): $N^2 \rightarrow E$
- We can estimate the parameters in linear time $O(E)$

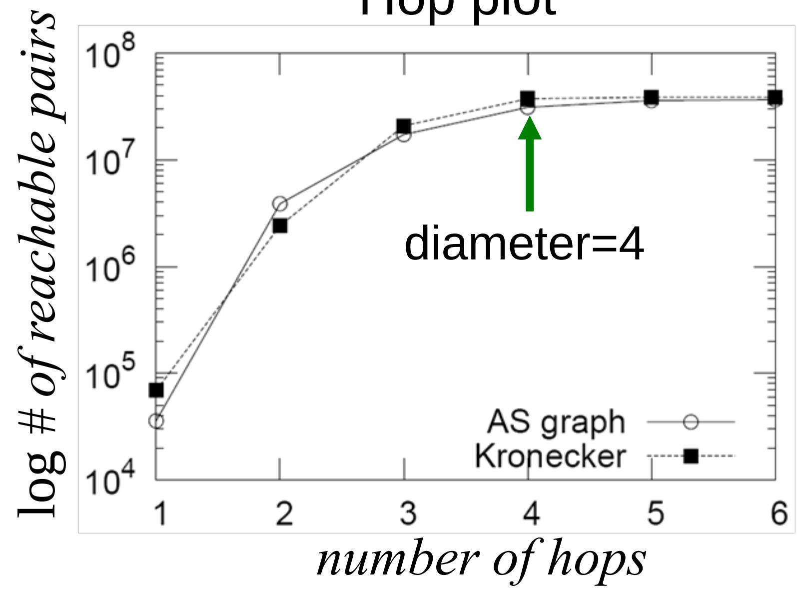
Model estimation: experiments

- Autonomous systems (internet): N=6500, E=26500
- Fitting takes 20 minutes
- AS graph is undirected and estimated parameters correspond to that

Degree distribution

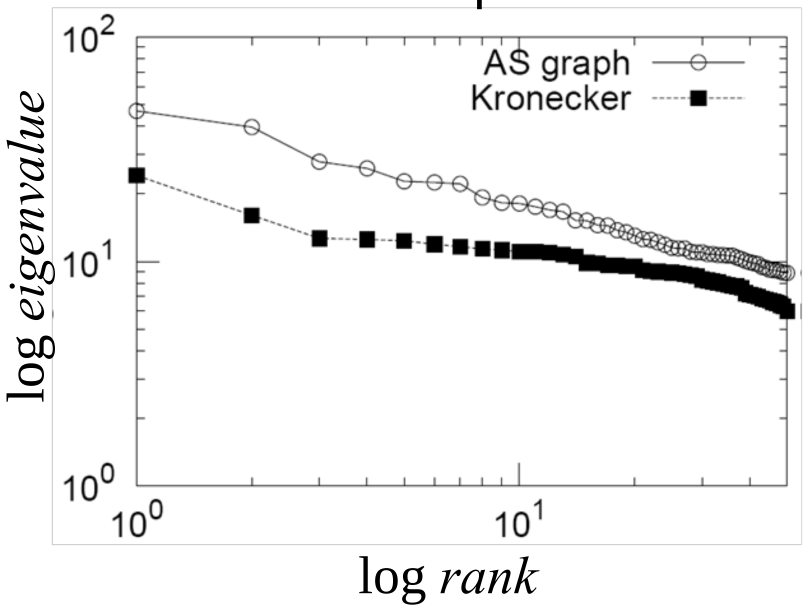


Hop plot

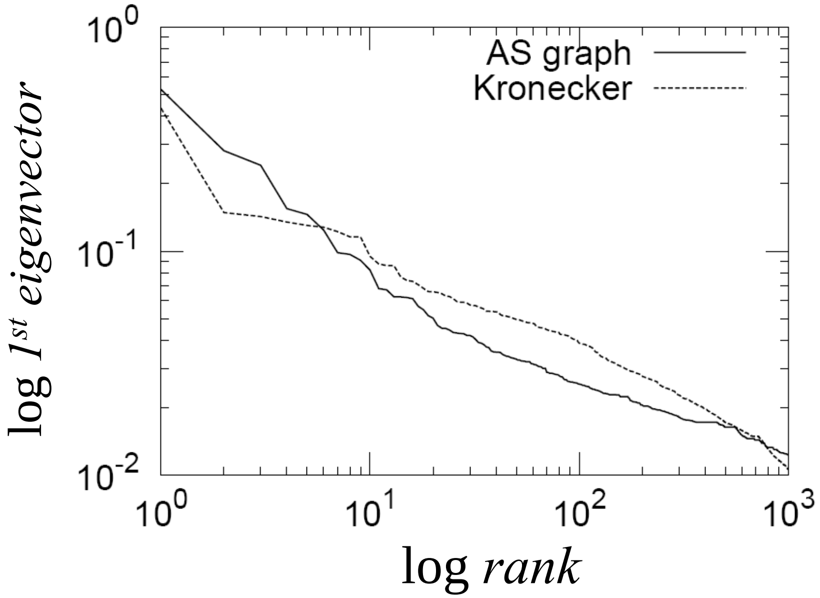


Model estimation: experiments

Scree plot



Network value

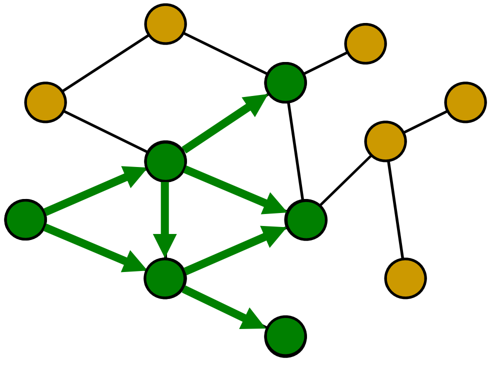


Completed work: Overview

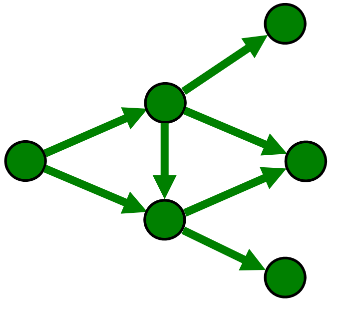
	S1: Dynamics of network evolution	S2: Dynamics of processes on networks
G1: Patterns	Densification Shrinking diameters	Cascade shape and size
G2: Models	Forest Fire Kronecker graphs	Cascade generation model
G3: Predictions	Estimating Kronecker parameters	Selecting nodes for detecting cascades

Information cascades

- Cascades are phenomena in which an idea becomes adopted due to influence by others



Social network



Cascade
(propagation graph)

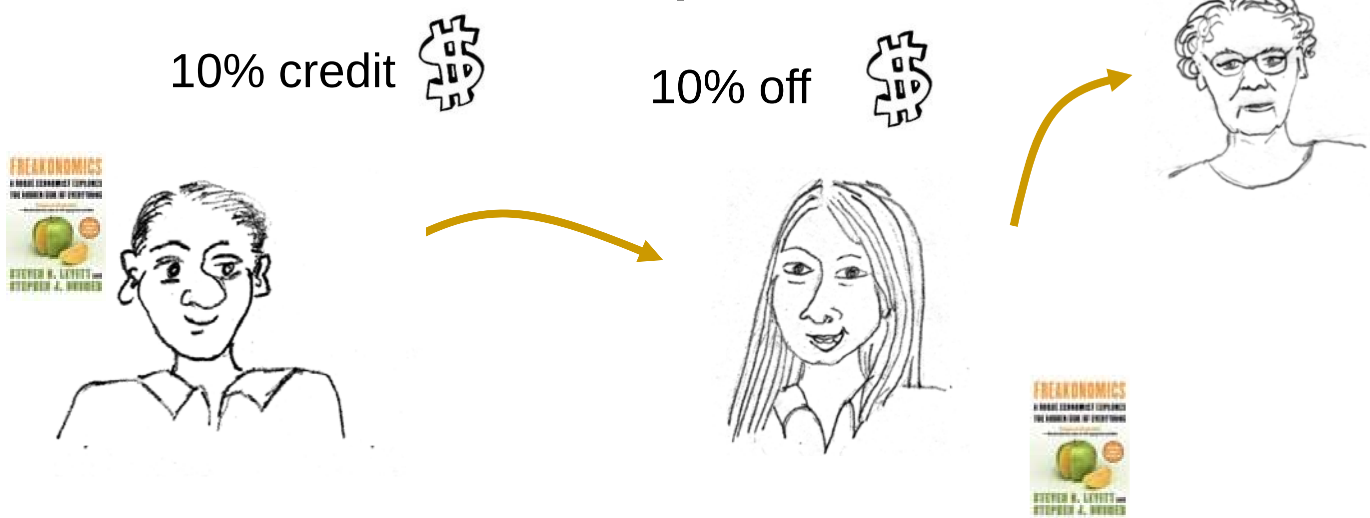
- We investigate cascade formation in
  - Viral marketing (Word of mouth)
  - Blogs

Cascades: Questions

- What kinds of cascades arise frequently in real life? Are they like trees, stars, or something else?
- What is the distribution of cascade sizes (exponential tail / heavy-tailed)?
- When is a person going to follow a recommendation?

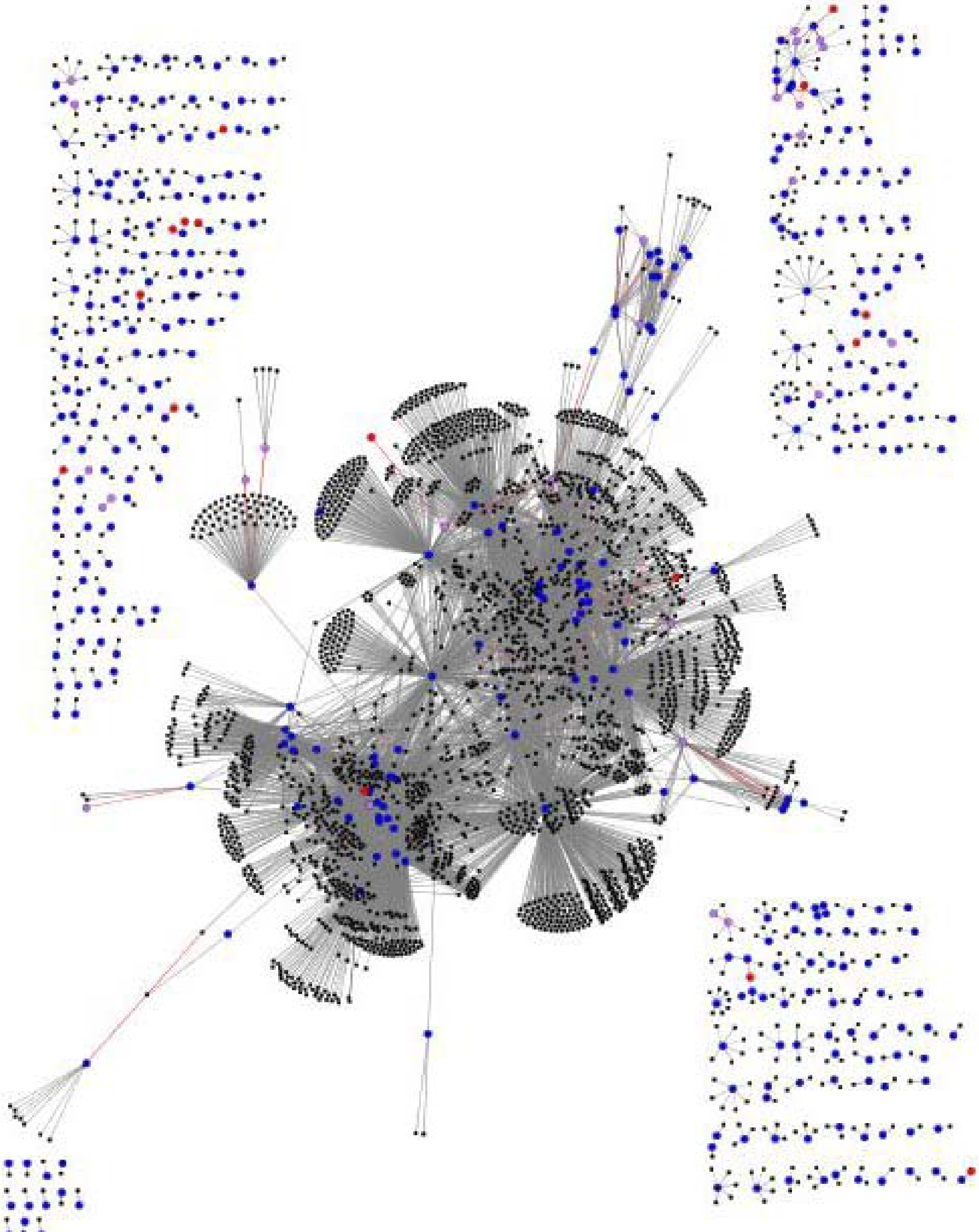
Cascades in viral marketing

- Senders and followers of recommendations receive discounts on products



  - Recommendations are made at time of purchase
  - Data: 3 million people, 16 million recommendations, 500k products (books, DVDs, videos, music)

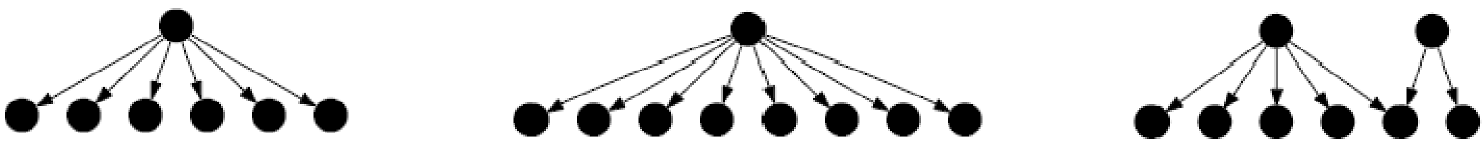
Product recommendation network



- ● purchase following a recommendation
- ● customer recommending a product
- • customer not buying a recommended product

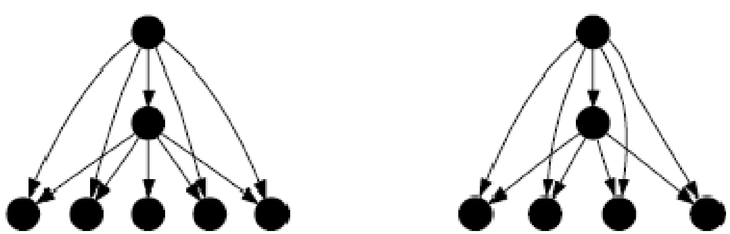
G1- Viral cascade shapes

- Stars (“no propagation”)



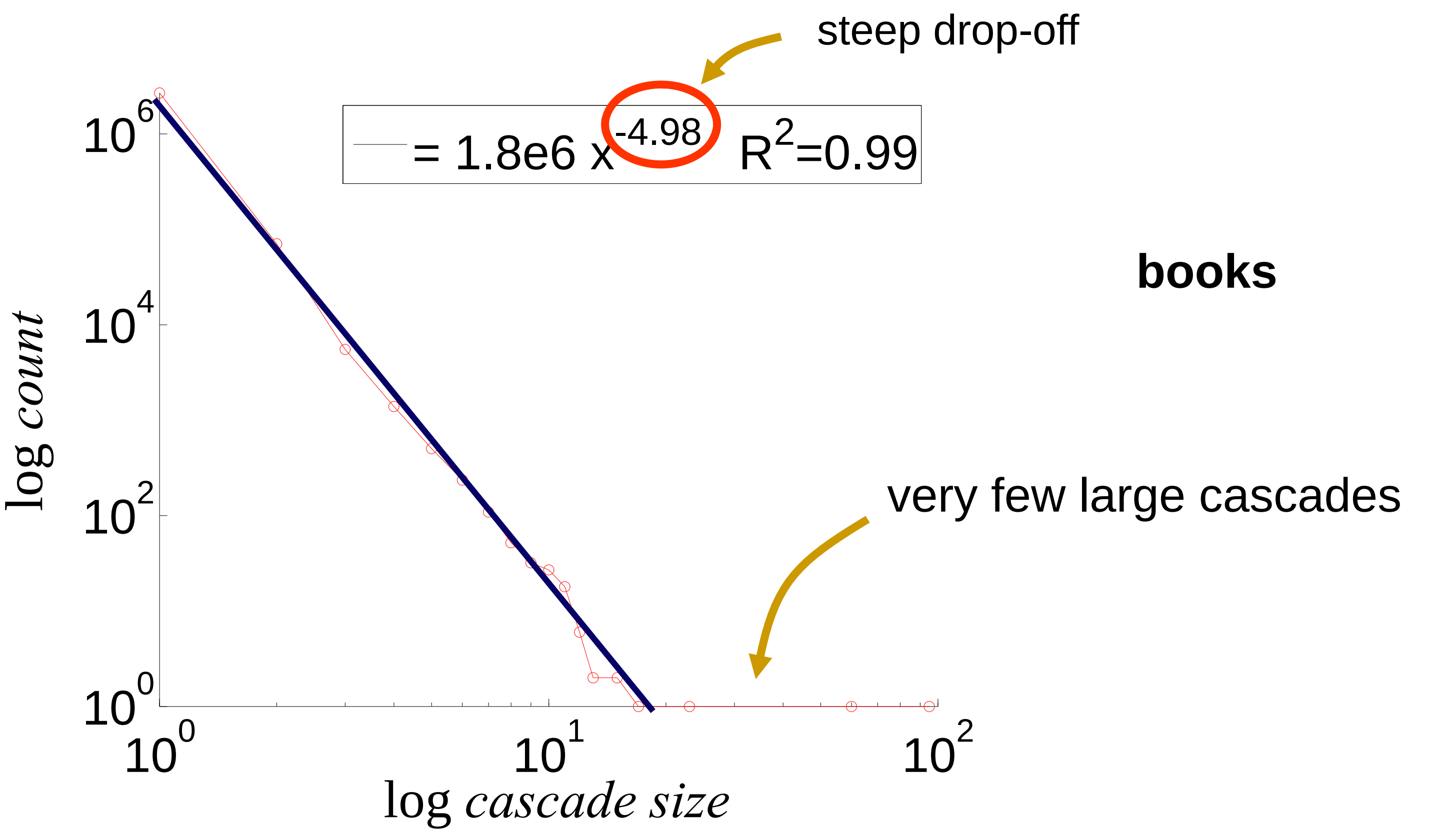
- Bipartite cores (“common friends”)

- Nodes having same friends

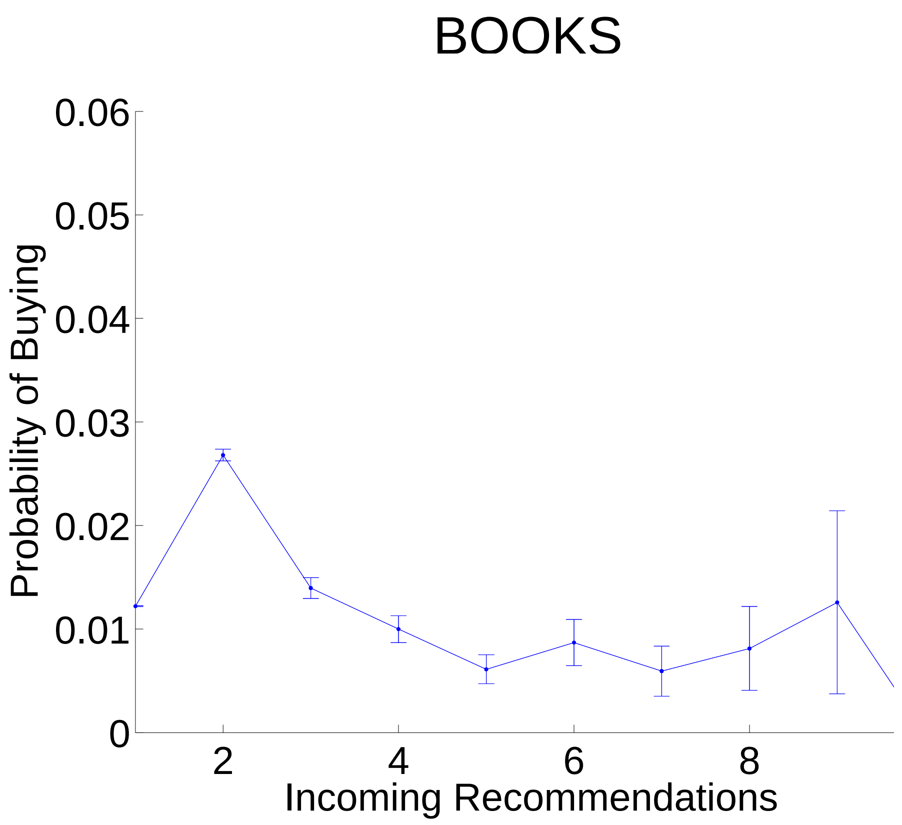


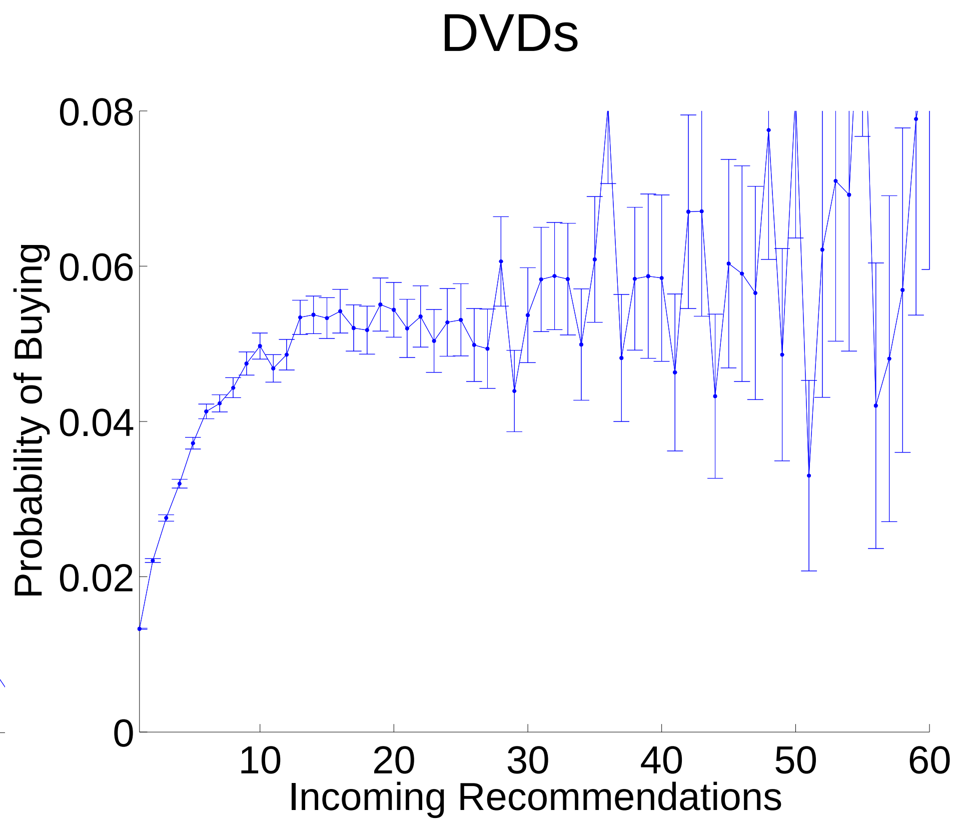
G1- Viral cascade sizes

- Count how many people are in a single cascade
- We observe a heavy tailed distribution which can not be explained by a simple branching process

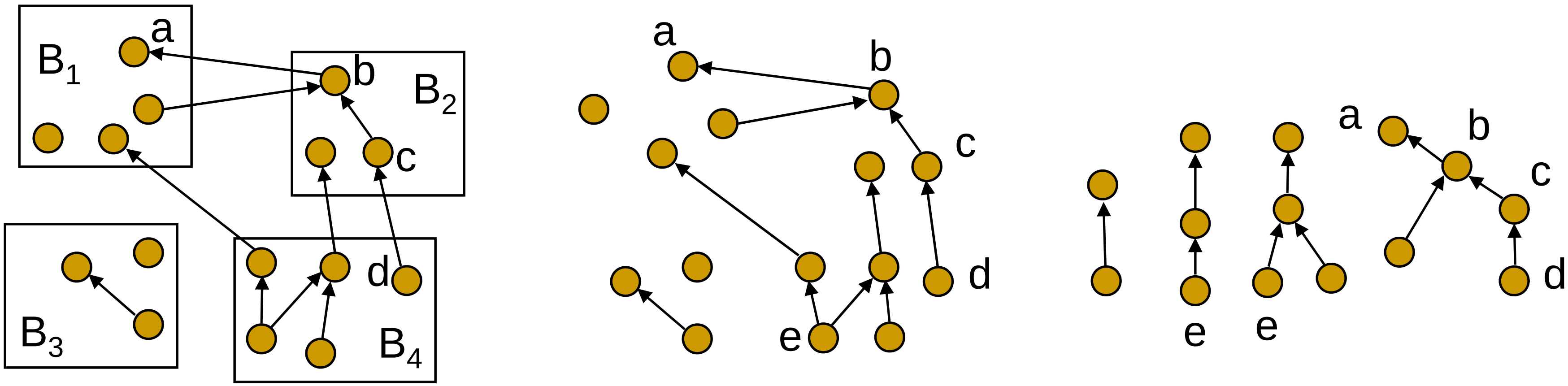


Does receiving more recommendations increase the likelihood of buying?





Cascades in the blogosphere



Blogosphere
blogs + posts

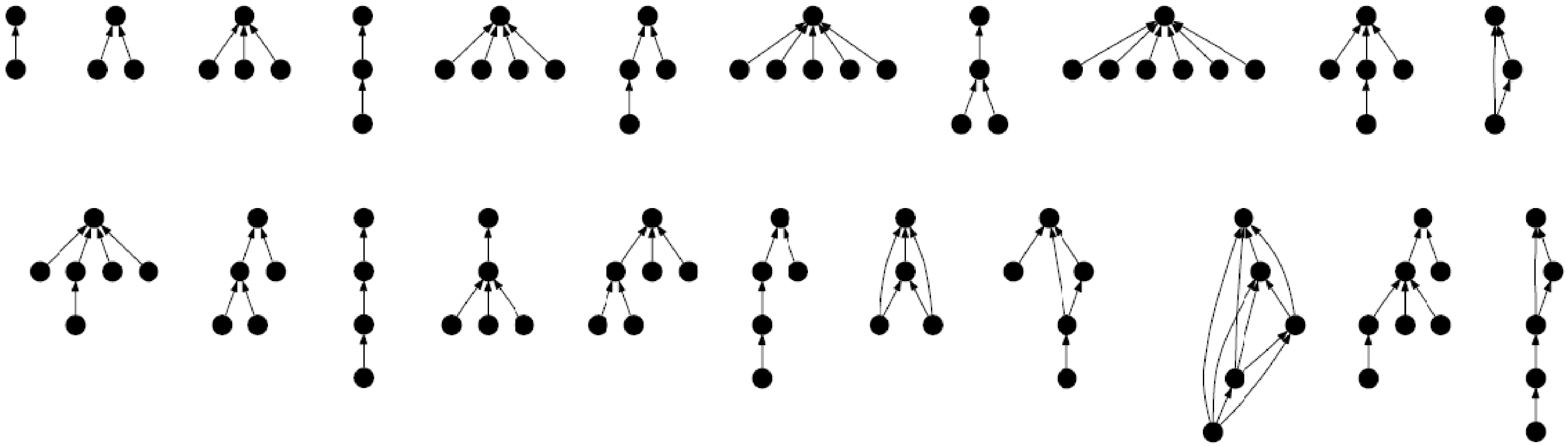
Post network
links among posts

Extracted
cascades

- Posts are time stamped
- We can identify cascades – graphs induced by a time ordered propagation of information

G1- Blog cascade shapes

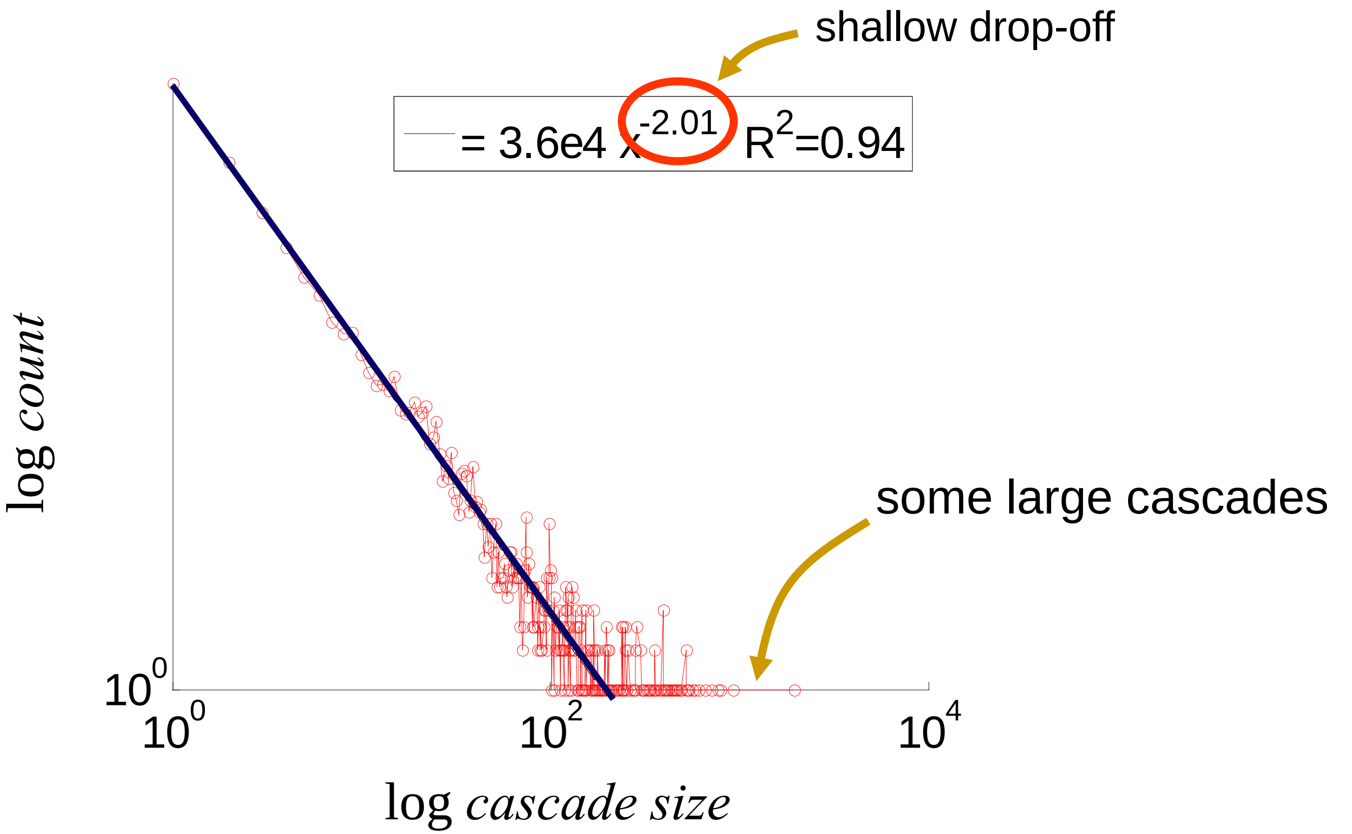
- Cascade shapes (ordered by frequency)



- Cascades are mainly stars
- Interesting relation between the cascade frequency and structure

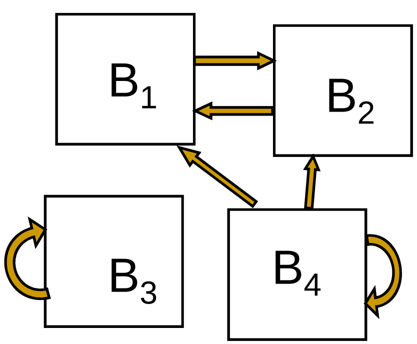
G1- Blog cascade size

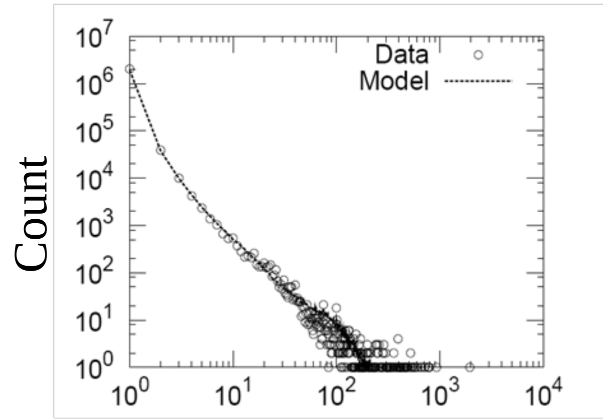
- Count how many posts participate in cascades
- Blog cascades tend to be larger than Viral Marketing cascades



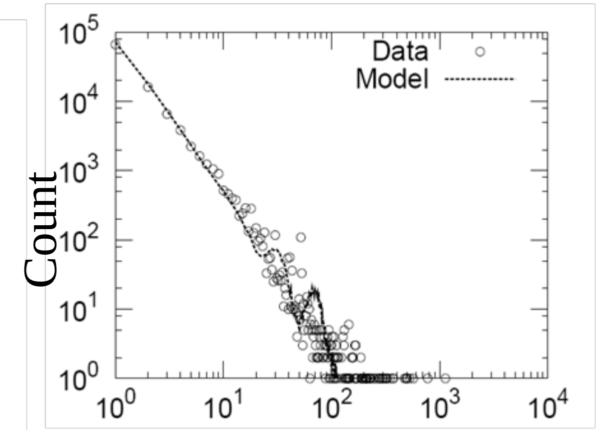
G2- Blog cascades: model

- Simple virus propagation type of model (SIS) generates similar cascades as found in real life

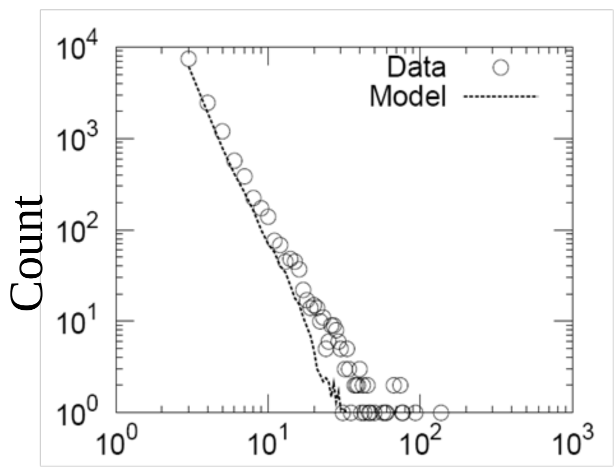




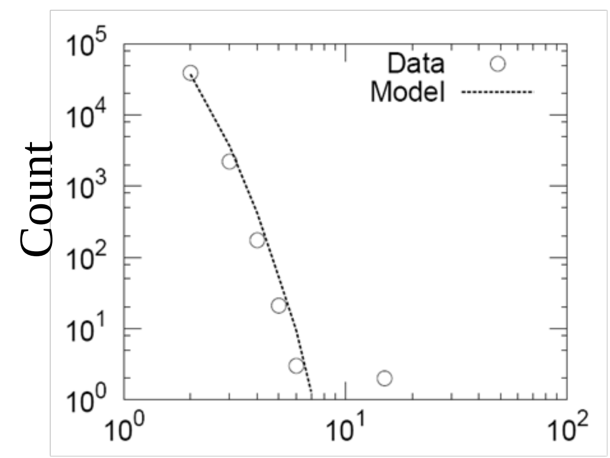
Cascade size



Cascade node in-degree



Size of star cascade



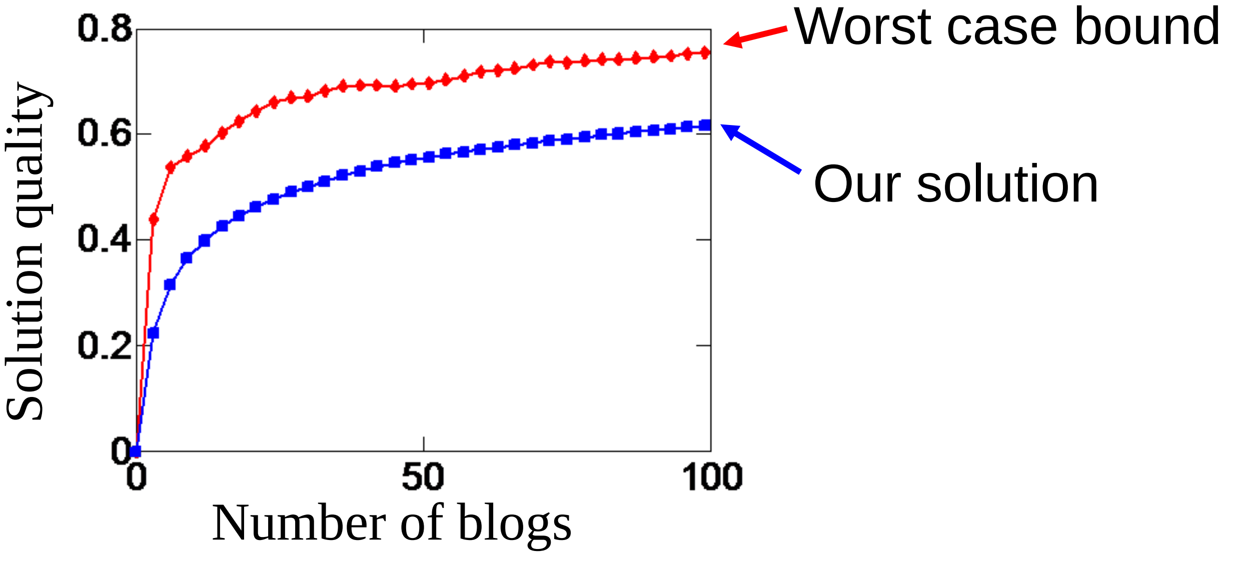
Size of chain cascade

G3- Node selection for cascade detection

- Observing cascades we want to select a set of nodes to quickly detect cascades
- Given a limited budget of attention/sensors
  - Which blogs should one read to be most up to date?
  - Where should we position monitoring stations to quickly detect disease outbreaks?

Node selection: algorithm

- Node selection is NP hard
- We exploit submodularity of objective functions to
  - develop scalable node selection algorithms
  - give performance guarantees



- In practice our solution is at most 5-15% from optimal

Outline

- Introduction
- Completed work
  - Network structure and evolution
  - Network cascades
- Proposed work
  - Large communication networks
  - Links and information cascades
  - Kronecker time evolving graphs
- Conclusion

Proposed work: Overview

	S1: Dynamics of network evolution	S2: Dynamics of processes on networks
G1: Patterns		1 Dynamics in communication networks
G2: Models		2 Models of link and cascade creation
G3: Predictions	3 Kronecker time evolving graphs	

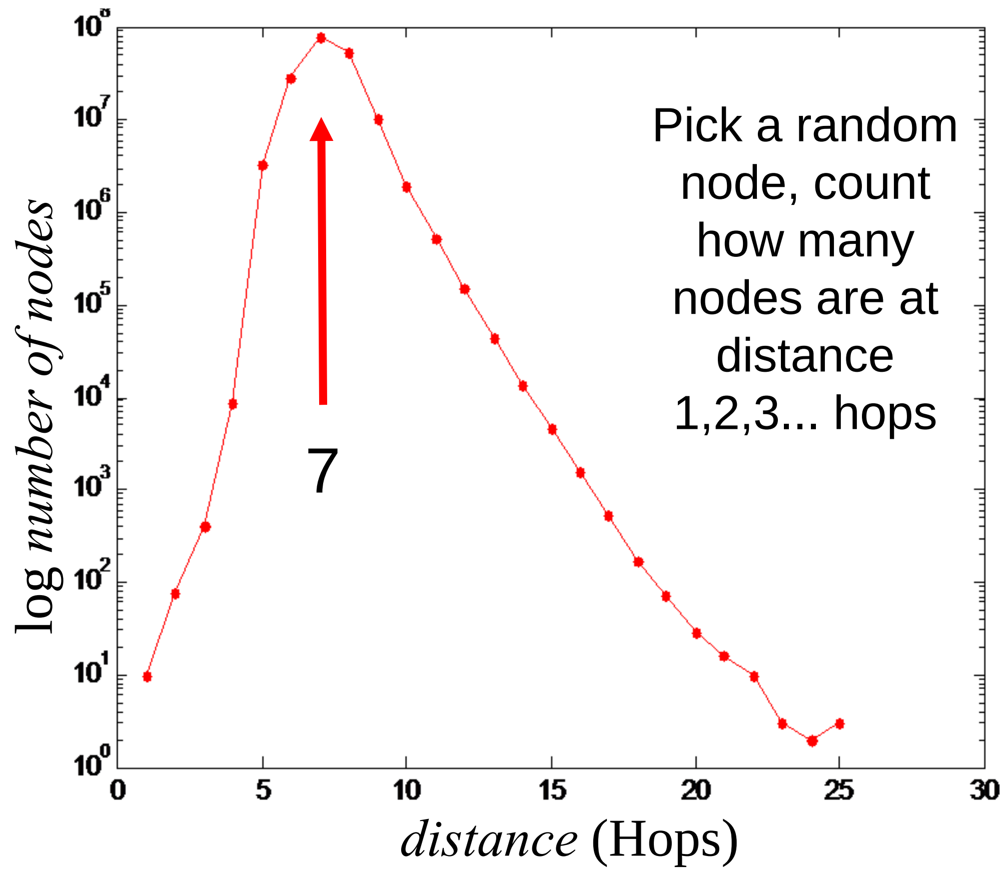
1 Proposed work: Communication networks

- Large communication network
  - 1 billion conversations per day, 3TB of data!
- How communication and network properties change with user demographics (age, location, sex, distance)
  - Test 6 degrees of separation
  - Examine transitivity in the network

1 Proposed work: Communication networks

- Preliminary experiment
  - Distribution of shortest path lengths
- Microsoft Messenger network
  - 200 million people
  - 1.3 billion edges
  - Edge if two people exchanged at least one message in one month period

MSN Messenger network

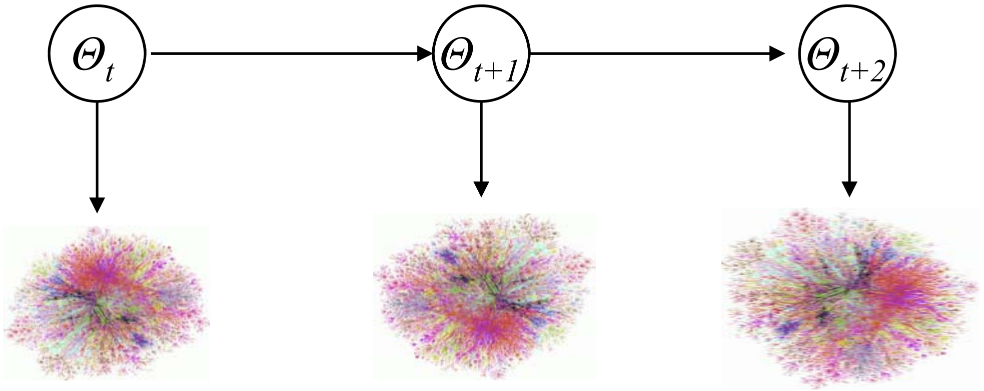


2 Proposed work: Links & cascades

- Given labeled nodes, how do links and cascades form?
- Propagation of information
  - Do blogs have particular cascading properties?
- Propagation of trust
  - Social network of professional acquaintances
    - 7 million people, 50 million edges
    - Rich temporal and network information
  - How do various factors (profession, education, location) influence link creation?
  - How do invitations propagate?

3 Proposed work: Kronecker graphs

- Graphs with weighted edges
  - Move beyond Bernoulli edge generation model
- Algorithms for estimating parameters of time evolving networks
  - Allow parameters to slowly evolve over time



Timeline

- May ‘07
  1 communication network
- Jun – Aug ‘07
  - research on on-line time evolving networks
- Sept– Dec ‘07
  2 Cascade formation and link prediction
- Jan – Apr ’08
  3 Kronecker time evolving graphs
- Apr – May ‘08
  - Write the thesis
- Jun ‘08
  - Thesis defense

References

- *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations,* by Jure Leskovec, Jon Kleinberg, Christos Faloutsos, ACM KDD 2005
- *Graph Evolution: Densification and Shrinking Diameters*, by Jure Leskovec, Jon Kleinberg and Christos Faloutsos, *ACM TKDD 2007*
- *Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication*, by Jure Leskovec, Deepay Chakrabarti, Jon Kleinberg and Christos Faloutsos, *PKDD 2005*
- *Scalable Modeling of Real Graphs using Kronecker Multiplication*, by Jure Leskovec and Christos Faloutsos, *ICML 2007*
- *The Dynamics of Viral Marketing,* by Jure Leskovec, Lada Adamic, Bernado Huberman, *ACM EC 2006*
- *Cost-effective outbreak detection in networks,* by Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, Natalie Glance, *in submission to KDD 2007*
- *Cascading behavior in large blog graphs,* by Jure Leskovec, Marry McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurst, SIAM DM 2007

Acknowledgements: Christos Faloutsos, Mary McGlohon, Jon Kleinberg, Zoubin Gharamani, Pall Melsted, Andreas Krause, Carlos Guestrin, Deepay Chakrabarti, Marko Grobelnik, Dunja Mladenic, Natasa Milic-Frayling, Lada Adamic, Bernardo Huberman, Eric Horvitz, Susan Dumais

Backup slides

1 Proposed work: Kronecker graphs

- Further analysis of Kronecker graphs
  - Prove properties of the diameter of Stochastic Kronecker Graphs
- Extend Kronecker to generate graphs with any number of nodes
  - Currently Kronecker can generate graphs with N^k nodes
  - Idea: expand only one row/column of current adjacency matrix

5 Proposed work: GraphGarden

- Publicly release a library for mining large graphs
  - Developed during our research
  - 40,000 lines of C++ code
- Components
  - Properties of static and evolving networks
  - Graph generation and model fitting
  - Graph sampling
  - Analysis of cascades
  - Node placement/selection

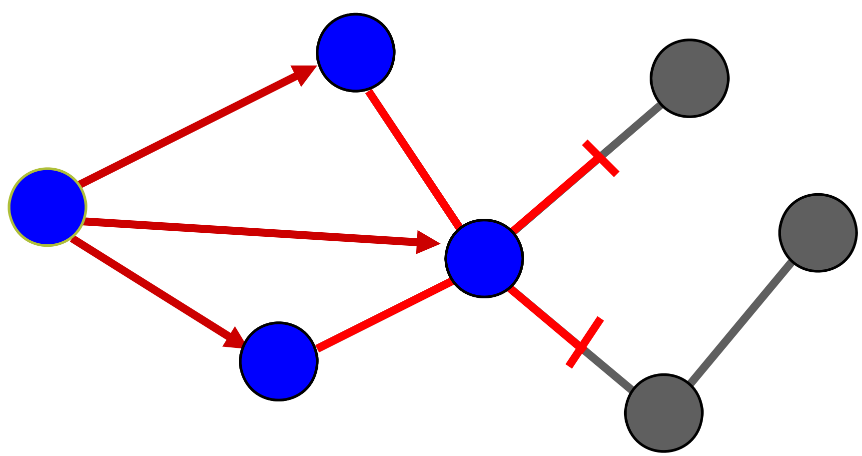
Proposed work

	Dynamics of network evolution	Dynamics of processes on networks
Structural properties		3 Dynamics in communication networks
Models	1 Analysis and extensions of Kronecker graphs	4 Models of link and cascade creation
Predictions	2 Kronecker time evolving graphs	

5 Release the graph mining toolkit

The model: Forest Fire Model

- Want to model graphs that density and have shrinking diameters
- Intuition:
  - How do we meet friends at a party?
  - How do we identify references when writing papers?

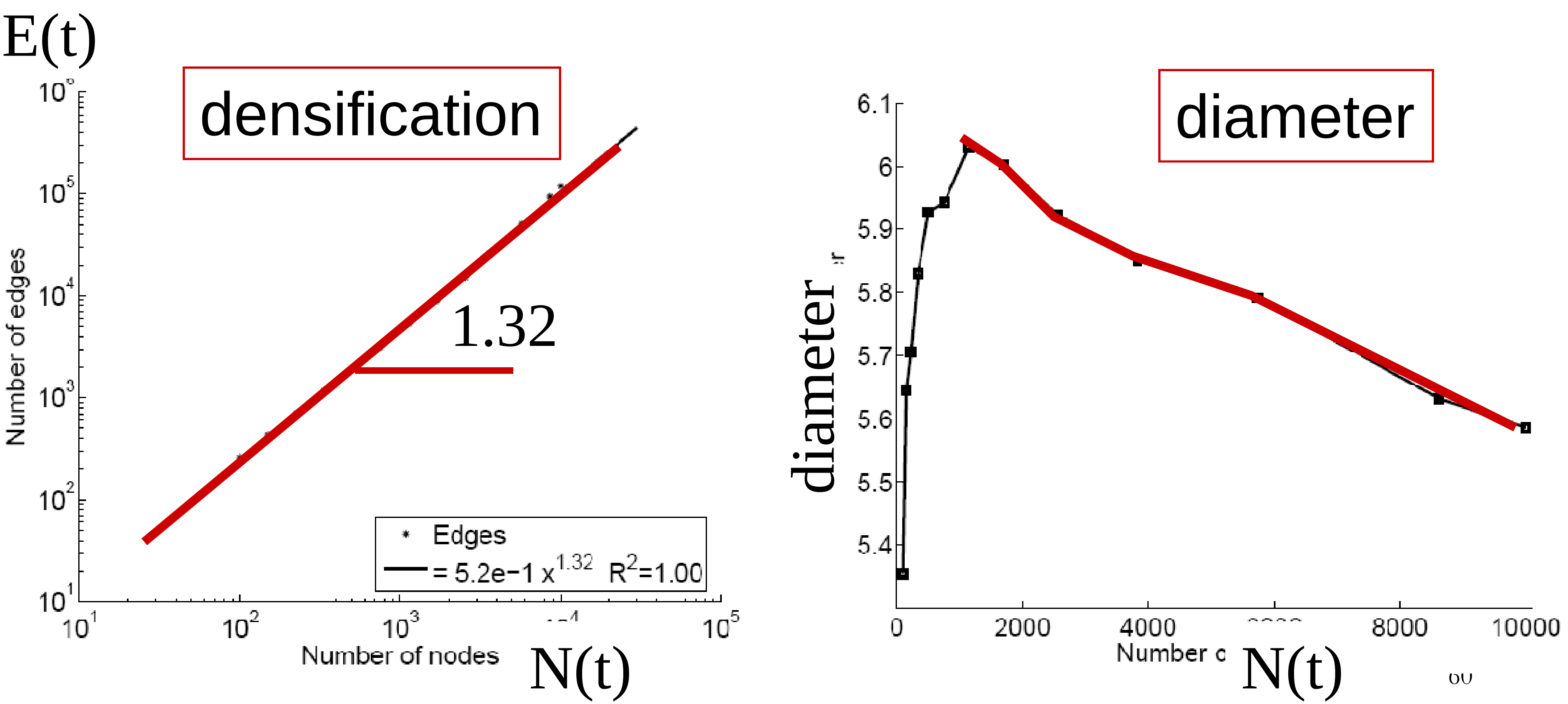


Properties of the Forest Fire

- Heavy-tailed in-degrees: “rich get richer”
  - Highly linked nodes can easily be reached
- Communities
  - Newcomer copies several of neighbors’ links
- Heavy-tailed out-degrees
  - Recursive nature provides chance for node to burn many edges
- Densification Power Law
  - Like in Community Guided Attachment
- Shrinking diameter
  - Densification helps but is not enough

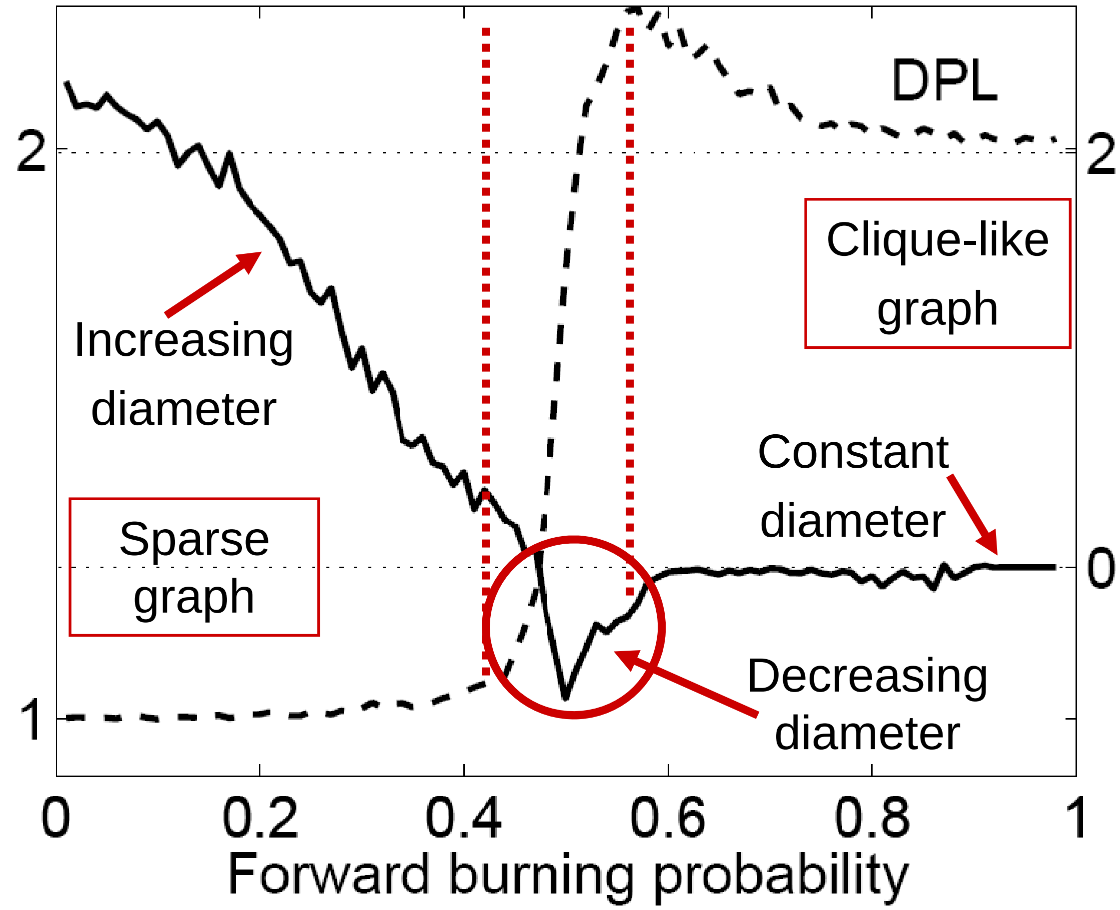
Forest Fire Model

- Forest Fire generates graphs that densify and have shrinking diameter



Forest Fire: Parameter Space

- Fix backward probability r and vary forward burning probability p
- We observe a sharp transition between sparse and clique-like graphs
- Sweet spot is very narrow



Kronecker product: Definition

- The Kronecker product of matrices *A* and *B* is given by

$$\underset{N\,x\,M}{\mathbf{C} = \mathbf{A}} \otimes \underset{K\,x\,L}{\mathbf{B}} \doteq \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}$$

*N*K x M*L*

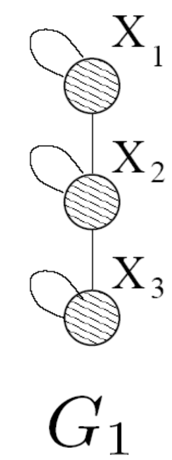
- We define a Kronecker product of two graphs as a Kronecker product of their adjacency matrices

Kronecker graphs

- We propose a growing sequence of graphs by iterating the Kronecker product

$$G_k = \underbrace{G_1 \otimes G_1 \otimes \ldots G_1}_{k\ times}$$

- Each Kronecker multiplication exponentially increases the size of the graph
- G_k has N_1^k nodes and E_1^k edges, so we get densification



Stochastic Kronecker graphs

- Create $N_1 \times N_1$ probability matrix P_1
- Compute the k^{th} Kronecker power P_k
- For each entry p_{uv} of P_k include an edge (u,v) with probability p_{uv}

0.5	0.2
0.1	0.3

P_1

Kronecker multiplication

Probability of edge p_{ij}

0.25	0.10	0.10	0.04
0.05	0.15	0.02	0.06
0.05	0.02	0.15	0.06
0.01	0.03	0.03	0.09

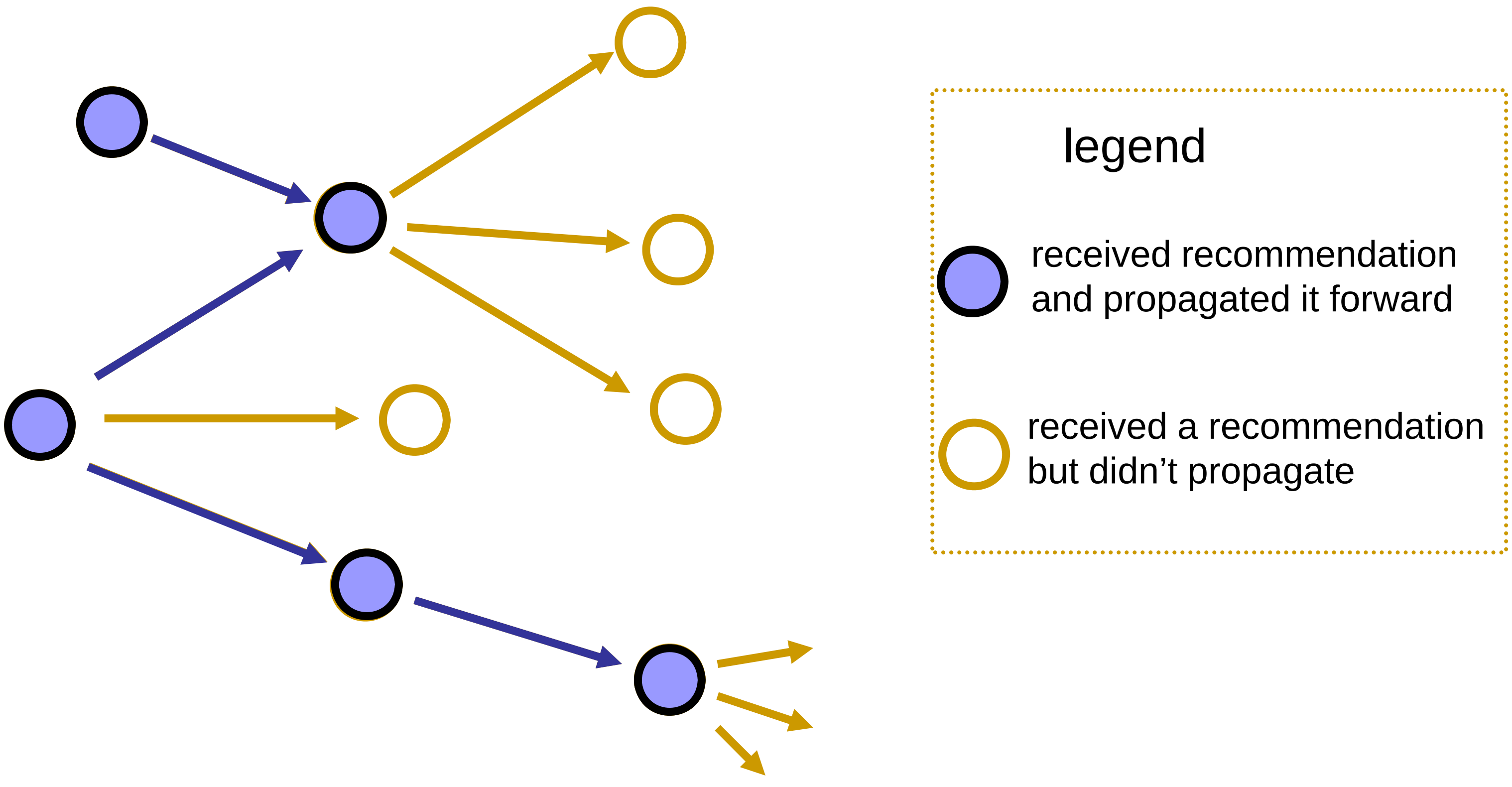
$P_2 = P_1 \otimes P_1$

Instance matrix K_2

flip biased coins

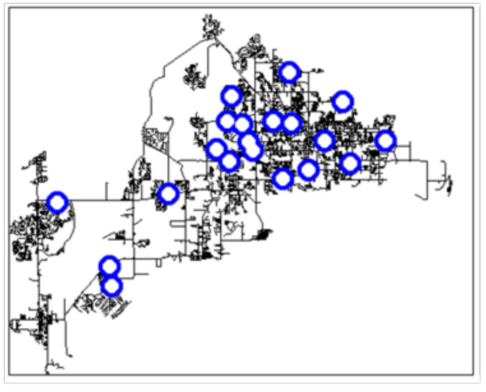
Cascade formation process

- Viral marketing
  - People purchase and send recommendations

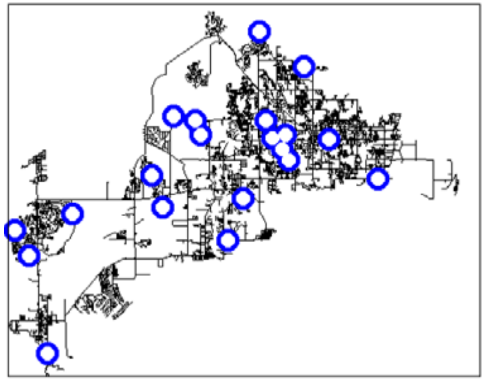


Node selection: example

- Water distribution network:
  - Different objective functions give different placements



Population affected



Detection likelihood