Dynamics of Real-world Networks

Thesis proposal

Jurij Leskovec
Machine Learning Department
Carnegie Mellon University

May 2, 2007

Thesis committee:
Christos Faloutsos, CMU
Avrim Blum, CMU
John Lafferty, CMU
Jon Kleinberg, Cornell University

Abstract

In our recent work we found very interesting and unintuitive patterns for time evolving networks, which change some of the basic assumptions that were made in the past. The main objective of observing the evolution patterns is to develop models that explain processes which govern the network evolution. Such models can then be fitted to real networks, and used to generate realistic graphs or give formal explanations about their properties. In addition, our work has a wide range of applications: we can spot anomalous graphs and outliers, design better graph sampling algorithms, forecast future graph structure and run simulations of network evolution.

Another important aspect of this research is the study of "local" patterns and structures of propagation in networks. We aim to identify building blocks of the networks and find the patterns of influence that these block have on information or virus propagation over the network. Our recent work included the study of the spread of influence in a large person-to-person product recommendation network and its effect on purchases. We also model the propagation of information on the blogosphere, and propose algorithms to efficiently find influential nodes in the network.

Further work will include three areas of research. We will continue investigating models for graph generation and evolution. Second, we will analyze large online communication networks and devise models on how user characteristics and geography relate to communication and network patterns. Third, we will extend the work on the propagation of influence in recommendation networks to blogs on the Web, studying how information spreads over the Web by finding influential blogs and analyzing their patterns of influence. We will also study how the local behavior affects the global structure of the network.

Contents

1 Introduction

The main interest of our research has been in understanding the structural properties and patterns in the evolution of large graphs and networks. What does a "normal" social network look like? How will it evolve over time? How can we spot "abnormal" interactions (*e.g.*, spam) in a time-evolving e-mail graph? How does information spread over the network? Answers to such questions are vital to a range of application areas from identification of illegal money-laundering rings, misconfigured routers on the Internet, to unexpected protein-protein interactions in a gene regulatory network.

Our proposed study of dynamics of large networks can be divided into two parts:

- The study of statistical properties and models that govern the generation and evolution of large real-world networks. We view the network as a big complex system, observe its static and temporal properties and patterns to design models that capture and help us understand the temporal and static patterns of real-world networks.

- The study of the network by starting from individual nodes and small communities. We are especially interested in modeling the spread of influence and information over the network and the substructures of the network, called *cascades*, that this process creates. We aim to find common and abnormal sub-network patterns and understand the propagation of influence, information, diseases and computer viruses over the network. Once we know the propagation patterns and structure, we devise algorithms for efficiently finding influential nodes.

In our work we focused on the way in which fundamental structural properties of networks vary with time. We found that two fundamental and commonly made assumptions about network evolution need to be reassessed. We discovered that networks densify as the network grows and that distances in the network shrink. As the existing graph generation models do not exhibit these types of behavior we developed three families of probabilistic generative models for graphs that capture these properties. The second part of our work focuses on the processes taking place in the networks. More specifically, we examined the structural and temporal properties of information propagation on large product recommendation and blog networks. We also created models of information propagation, and developed scalable algorithms for finding influential nodes in the network.

Our studies involve large real-world datasets with millions of nodes and edges. Working with such datasets is important in order to understand and take into account performance and scalability issues and to discover patterns that may become apparent only in massive datasets.

1.1 Motivation

Traditionally small networks were analyzed from a "node centric" point of view where researchers wanted to answer questions about behavior and properties of particular nodes in the network. Though, such models are very expressive, they often fail to scale to large networks with millions of nodes and edges. Moreover, many times we need to work with a large network for a structural property of the network to emerge, thus the focus moves to the study of structural properties of the network as a whole.

1.1.1 Network structure and models

Ultimately we search for interesting measures that let us characterize the network structure and the processes spreading over the networks. Then we design models and algorithms that take advantage of the identified structural network properties.

The focus of analyzing and modeling the structure of large networks aims to do the following three things:

(1) *What are interesting statistical properties of network structure?* The aim is to find statistical properties, such as path lengths and degree distributions, that characterize the structure and behavior of networks, and suggest appropriate ways to measure these properties.

(2) *What is a good model that helps us understand these properties?* We aim in creating models of networks that can help us to understand the meaning of the statistical properties of networks. How they come to be as they are, and how they interact with one another?

(3) *Predict behavior of networks based on measured structural properties and local rules governing individual nodes?* How, for example, will Internet structure evolve and how does the network structure affect traffic on the Internet or performance of a web crawler?

1.1.2 Cascading behavior in large networks

The second part of the thesis deals with information propagation in the large networks. The social network of interactions among a group of individuals plays a fundamental role in the spread of information, ideas, and influence. Such effects have been observed in many cases, when an idea or action gains sudden widespread popularity through word-of-mouth or "viral marketing" effects. To take a recent example from the technology domain, free e-mail services such as Microsoft's Hotmail and later Google's Gmail achieved wide usage largely through referrals, rather than direct advertising.

We would like to understand how the structure of the network affects the spread of information, influence and viruses over the network. We monitor the spread of information on the blogosphere or recommendations in a product recommendation network. For example, when studying information propagation on the blogosphere, we ask what are the typical structural patterns of information propagation? How deep or wide are the propagation graphs (also called *cascades*)? How fast is the information spreading? We also aim in creating models and algorithms that help us predict future and identify influential nodes, *e.g.*, given a fixed budget of attention, which blogs should we read to be most up to date on the news? Or similarly, in a big water distribution network, where shall we position the sensors to detect disease outbreaks as quickly as possible?

1.2 Applications, consequences and impact

Accurate properties of network growth, information propagation, and the models supporting them, have several possible application and consequences. Patterns give us ways for understanding and building models, and models help us to reason, monitor and predict features of the network in the future.

- *Models and parameters:* Generative models and their parameters give us insight into graph formation process. Intuitions developed by the models are useful in understanding the network generation processes and reasoning about the structure of the networks in general.

- *Graph generation:* Our methods form a means of assessing the quality of graph generators. Synthetic graphs are important for "what if" scenarios where we need to extrapolate and simulate graph growth and evolution, since real graphs may be impossible to collect and track (like, *e.g.*, a very large friendship graph between people). Synthetic graphs can then be used for simulations and evaluation of algorithms, *e.g.*, simulations of new network routing protocols, virus propagation, etc.

- *Graph sampling:* Large real-world graphs are becoming increasingly available, with sizes ranging from the millions to billions of nodes. There are many algorithms for computing interesting graph properties (shortest paths, centrality, betweenness, etc.), but most of these algorithms become impractical for large graphs. Thus sampling is essential — but sampling from a graph is a non-trivial problem. Densification laws can help discard bad sampling methods, by providing a means to reject poorly sampled subgraphs.

- *Extrapolations:* For several real graphs, we have a lot of snapshots of their past. What can we say about their future? Our results help form a basis for validating scenarios for graph evolution.

- *Abnormality detection and computer network management:* In many network settings, "normal" behavior will produce subgraphs that obey densification laws (with a predictable exponent) and other properties of network growth. If we detect activity producing structures that deviate significantly from the normal patterns, we can flag them as abnormalities; this can potentially help with the detection of, e.g., fraud, spam, or distributed denial of service (DDoS) attacks.

- *Graph compression:* In many cases one would want to efficiently describe the graph. This can be done by compressing the graph by just storing the set of model parameters, and then the deviations between the real and the synthetic graph.

- *Anonymization:* Suppose that the real graph can not be publicized, like, *e.g.*, corporate e-mail network or customer-product sales in a recommendation system. Yet, we would like to share our network. One can use our findings and models we developed as a way to generate a similar synthetic network.

- *Network cascades:* Understanding cascade formation helps to explain the propagation of information and viruses over the network. This allows for more accurate models of virus propagation, which can be used in epidemiology for simulations.

Our published work as it maps to the chapters of the thesis proposal

- Section 3.1 and Section 3.2
  - [**Paper A**] Leskovec, J., Kleinberg, J. M., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD).*
  - Leskovec, J., Kleinberg, J. M., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: ACM SIGKDD conference on Knowledge discovery in data mining.*
- Section 3.3
  - [**Paper B**] Leskovec, J., Chakrabarti, D., Kleinberg, J. M., and Faloutsos, C. (2005). Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication. In *PKDD '05: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases.*
  - [**Paper C**] Leskovec, J. and Faloutsos, C. (2007). Scalable modeling of real graphs using Kronecker multiplication. In *ICML '07: International Conference on Machine Learning.*
- Section 4.1
  - [**Paper D**] Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB).*
  - Leskovec, J., Singh, A., and Kleinberg, J. M. (2006). Patterns of influence in a recommendation network. In *PAKDD '06: Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining.*
  - Leskovec, J., Adamic, L. A., and Huberman, B. A. (2006). The dynamics of viral marketing. In *EC '06: 7th ACM conference on Electronic commerce.*
- Section 4.2
  - [**Paper E**] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. S., and Hurst, M. (2007). Cascading behavior in large blog graphs. In *SDM '07: SIAM Conference on Data Mining.*
- Section 4.3
  - [**Paper F**] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). Cost-effective outbreak detection in networks. *Submitted to KDD '07.*

- *Outbreak detection:* Our work on cascades also gives us the means to study, for example, which nodes to inoculate to prevent a virus from spreading through the network, or where to place sensors in a water distribution network to quickly detect disease outbreaks.

We applied our findings and models in the following applications. Structural patterns of networks help us design better graph sampling techniques (Leskovec and Faloutsos, 2006).

Exploiting the graph structure also helps us improve on various machine learning tasks. For example, in a web search application we recently showed (Leskovec et al., 2007b) that by exploiting the graph structure of the Web one can predict the quality of the obtained search results and the amount of spam web-pages in the search results. Similarly, measurements and models of information and virus propagation give us means to determine conditions under which the information will die-out or remain in the network (Chakrabarti et al., 2007), and to develop algorithms for selecting nodes to early detect information or virus epidemics in networks (Leskovec et al., 2007d).

In the following sections we briefly present some of our recent work on the dynamics of networks themselves and processes taking place on them. First, we survey the related work on statistical properties of networks, generative models and network cascades in section 2. Our work on properties and models of network evolution is presented in section 3. Section 4 discusses the results dynamics of processes taking place in the networks. We present the plan of future research section 5, and conclude in section 6.

2 Survey

Next, we briefly survey the related work. First, we focus on properties of static networks and continue with surveying the work on explanatory models. Last, we introduce the work on cascades and information propagation in networks.

2.1 Properties of networks

Networks are composed of nodes and edges connecting them. Examples of networks include the Internet, World Wide Web, social networks of acquaintance, collaboration or other connections between individuals, organizational networks, metabolic networks, language networks, food webs, distribution networks such as water distribution networks, blood vessels or postal delivery routes, networks of citations between papers, software networks where edges represent dependencies or function calls.

Research over the past few years has identified classes of properties that can be found in many real-world networks from various domains. While many patterns have been discovered, two of the principal ones are heavy-tailed degree distributions and small diameters.

Degree distribution: A distribution is a Power-law if it has a PDF (probability density function) of the form $p(x) \propto x^{\gamma}$, where $p(x)$ is the probability to encounter value x and γ is the exponent of the power law. In log-log scales, such a PDF gives a straight line with slope γ. For $\gamma < -1$, we can show that the Complementary Cumulative Distribution Function (CCDF) is also a power law with slope $\gamma + 1$, and so is the rank-frequency plot pioneered by Zipf (Zipf, 1949), with slope $1/(1 + \gamma)$. For $\gamma = -2$ we have the standard Zipf distribution, and for other values of γ we have the generalized Zipf distribution.

The degree-distribution of a graph is a power law if the number of nodes c_k with degree k is given by $c_k \propto k^{-\gamma} \quad (\gamma > 0)$ where γ is called the power-law exponent. Power laws have been found in the Internet (Faloutsos et al., 1999), the Web (Kleinberg et al., 1999, Broder et al., 2000, Albert and Barabasi, 1999, Huberman and Adamic, 1999, Kumar et al., 1999), citation graphs (Redner, 1998), click-stream data (Bi et al., 2001), online social networks (Chakrabarti et al., 2004) and many others. Deviations from the power-law pattern have been noticed (Pennock et al., 2002), which can be explained by the "DGX" distribution (Bi et al., 2001).

Small diameter: Most real-world graphs exhibit relatively small diameter (the "small-world" phenomenon): A graph has diameter d if every pair of nodes can be connected by a path of length at most d. The diameter d is susceptible to outliers. Thus, a more robust measure of the pairwise distances between nodes of a graph is the *effective diameter* (Tauro et al., 2001). This is defined as the minimum number of hops in which some fraction (or quantile q, say $q = 90\%$) of all connected pairs of nodes can reach each other. The effective diameter has been found to be small for large real-world graphs, like Internet, Web, and social networks (Albert and Barabási, 2002, Milgram, 1967, Albert et al., 1999, Bollobas and Riordan, 2004, Broder et al., 2000, Chung and Lu, 2002, Watts and Strogatz, 1998)).

Scree plot: This is a plot of the eigenvalues (or singular values) of the adjacency matrix of the graph, versus their rank, using a log-log scale. The scree plot is also often found to approximately obey a power law (Dorogovtsev et al., 2002). The distribution of eigenvector components (indicators of "network value") has also been found to be skewed (Chakrabarti et al., 2004).

Clustering coefficient: This is a measure of transitivity in networks (Watts and Strogatz, 1998), i.e., friend of a friend is also my friend. In many networks it is found that if node u is connected to v and v is further connected to w then there is a higher probability that node u is connected to w. In terms of network topology, transitivity means the presence of a heightened number of triangles in the network, i.e. sets of three fully connected nodes. Clustering coefficient C_k of a vertex of degree k is defined as follows. Let node v have k neighbors; then at most $k(k-1)/2$ edges can exist between them. Let C_v denote the fraction of these allowable edges that actually exist. Then C_k is defined as the average C_v over all nodes v of degree k, and the global clustering coefficient C is the average C_v over all nodes v. Clustering coefficient in real networks is significantly higher than for random networks. It has also been found that in scale-free and real networks clustering C_k scales as k^{-1} (Dorogovtsev et al., 2002, Ravasz and Barabasi, 2003).

Community structure: Many networks and most social networks show some community structure (Wasserman et al., 1994, Girvan and Newman, 2002). Intuitively this means that there are groups of nodes that have a high density of connections within them, and a lower density of connections between the groups. Many times it is found that the communities observe a recursive structure, where bigger communities can further be split into smaller and smaller communities.

Apart from these, several other patterns have been found, including the "resilience" (Albert and Barabási, 2002, Palmer et al., 2002), which shows that real-networks are resilient to random node attacks; Other properties are "stress" (Chakrabarti et al., 2004), network navigation (Kleinberg, 1999, Watts et al., 2002), and many more. We point the reader to (Mitzenmacher, 2004, Newman, 2003, Li et al., 2005) for overviews of this area.

2.2 Explanatory models

In parallel with empirical studies of large networks, there has been considerable work on probabilistic models for graph generation.

The earliest probabilistic generative model for graphs was a random graph model, where each pair of nodes has an identical, independent probability of being joined by an edge (Erdős and Rényi, 1960). The study of this model has led to a rich mathematical theory; however, this generator produces graphs that fail to match real-world networks in a number of respects (e.g., it does not produce heavy-tailed degree distributions).

The discovery of degree power laws led to the development of random graph models that exhibited such degree distributions, including the family of models based on *preferential attachment* (Albert and Barabasi, 1999, Cooper and Frieze, 2003, Aiello et al., 2000): new nodes join the graph at each time step, and preferentially connect to existing nodes with high degree (the "rich get richer") (Simon, 1955). This simple behavior leads to power-law tails and to low diameters. The diameter in this model grows slowly, i.e. logarithmically, with the number of nodes, which violates the "shrinking diameter" property we describe later.

Similar in spirit are the *copying model* (Kleinberg et al., 1999, Kumar et al., 2000), the related *growing network with copying* model (Krapivsky and Redner, 2005), and models based on random walks (Blum et al., 2006) and recursive search (Vazquez, 2001) for generating networks. The common theme among these models is that a node joins the network by uniformly at random choosing node u and then either link ing to u's neighbors, start a random walk or breath first search type of procedure to create links to nodes in u's vicinity.

Another family of graph-generation methods strives for small diameter, like the *small-world* generator (Watts and Strogatz, 1998) and the Waxman generator (Waxman, 1988). A third family of methods show that heavy tails emerge if nodes try to optimize their connectivity under resource constraints (Carlson and Doyle, 1999, Fabrikant et al., 2002). Recent work of (Chakrabarti and Faloutsos, 2006) gives a survey of the structural properties and statistics of real world graphs and the underlying generative models for graphs.

2.3 Cascades in networks

Information cascades are phenomena in which an action or idea becomes widely adopted due to influence by others (Bikhchandani et al., 1992). Cascades are also known as "fads" or "resonance." Cascades have been studied for many years by sociologists concerned with the *diffusion of innovation* (Rogers, 1995); more recently, researchers in several fields have investigated cascades for the purpose of selecting trendsetters for viral marketing (Domingos and Richardson, 2001), finding inoculation targets in epidemiology (Newman et al., 2002), and explaining trends in blogspace (Kumar et al., 2003). Despite much empirical work in the social sciences on datasets of moderate size, the difficulty in obtaining data has limited the extent of analysis on very large-scale, complete datasets representing cascades. We look at the patterns of influence in a large-scale, real recommendation network and examine the topological structure of cascades.

Most of the previous research on the flow of information and influence through the networks has been done in the context of epidemiology and the spread of diseases over the network (Bailey, 1975, Anderson and May, 2002). Classical disease propagation models are based on the stages of a disease in a host: a person is first *susceptible* to a disease, then if she is exposed to an infectious contact she can become *infected* and thus *infectious.* After the disease ceases the person is *recovered* or *removed.* Person is then *immune* for some period.

The immunity can also wear off and the person becomes again susceptible. Thus SIR (susceptible – infected – recovered) models diseases where a recovered person never again becomes susceptible, while SIRS (SIS, susceptible – infected – (recovered) – susceptible) models population in which recovered host can become susceptible again. Given a network and a set of infected nodes the *epidemic threshold* is studied, *i.e.*, conditions under which the disease will either dominate or die out.

Diffusion models that try to model the process of adoption of an idea or a product can generally be divided into two groups:

- *Threshold model* (Granovetter, 1978) where each node in the network has a threshold $t \in [0, 1]$, typically drawn from some probability distribution. We also assign *connection weights* $w_{u,v}$ on the edges of the network. A node adopts the behavior if a sum of the connection weights of its neighbors that already adopted the behavior (purchased a product in our case) is greater than the threshold: $t \leq \sum_{\text{adopters}(u)} w_{u,v}$.

- *Independent cascade model* (Goldenberg et al., 2001) where whenever a neighbor v of node u adopts, then node u also adopts with probability $p_{u,v}$. In other words, every time a neighbor of u purchases a product, there is a chance that u will decide to purchase as well.

While these models address the question of how influence spreads in a network, they are based on *assumed* rather than *measured* influence effects. In contrast, our study tracks the actual diffusion of recommendations through email, allowing us to quantify the importance of factors such as the presence of highly connected individuals, or the effect of receiving recommendations from multiple contacts. Compared to previous empirical studies which tracked the adoption of a single innovation or product, our data encompasses over half a million different products, allowing us to model a product's suitability for viral marketing in terms of both the properties of the network and the product itself.

2.3.1 Information cascades in blogosphere

Most work on extracting cascades has been done in the blog domain (Adamic and Glance, 2005, Adar and Adamic, 2005, Gruhl et al., 2004). The authors in this domain noted that, while information propagates between blogs, examples of genuine cascading behavior appeared relatively rarely. This is possibly due to bias in the web-crawling and text analysis techniques used to collect pages and infer relationships. In our dataset, all the recommendations are stored as database transactions, and we know that no records are missing. Associated with each recommendation is the product involved, and the time the recommendation was made. Studies of blogspace either spend a lot of effort mining topics from posts (Adar and Adamic, 2005, Gruhl et al., 2004) or consider only the properties of blogspace as a graph of unlabeled URLs (Adamic and Glance, 2005).

There are several potential models to capture the structure of the blogosphere. Work on information diffusion based on topics (Gruhl et al., 2004) showed that for some topics, their popularity remains constant in time ("chatter") while for other topics the popularity is more volatile ("spikes"). (Kumar et al., 2003) analyze community-level behavior as inferred from blog-rolls – permanent links between "friend" blogs. In their extension (Kumar et al.,

2006) performed analysis of several topological properties of link graphs in communities, finding that much behavior was characterized by “stars”.

2.3.2 Cascades in viral marketing

Viral marketing can be thought of as a diffusion of information about the product and its adoption over the network. Primarily in social sciences there is a long history of research on the influence of social networks on innovation and product diffusion. However, such studies have been typically limited to small networks and typically a single product or service. For example, (Brown and Reingen, 1987) interviewed the families of students being instructed by three piano teachers, in order to find out the network of referrals. They found that strong ties, those between family or friends, were more likely to be activated for information flow and were also more influential than weak ties (Granovetter, 1973) between acquaintances.

In the context of the internet, word-of-mouth advertising is not restricted to pairwise or small-group interactions between individuals. Rather, customers can share their experiences and opinions regarding a product with everyone. Quantitative marketing techniques have been proposed (Montgomery, 2001) to describe product information flow online, and the rating of products and merchants has been shown to effect the likelihood of an item being bought (Resnick and Zeckhauser, 2002, Chevalier and Mayzlin, 2006). More sophisticated online recommendation systems allow users to rate others' reviews, or directly rate other reviewers to implicitly form a trusted reviewer network that may have very little overlap with a person's actual social circle. (Richardson and Domingos, 2002) used Epinions' trusted reviewer network to construct an algorithm to maximize viral marketing efficiency assuming that individuals' probability of purchasing a product depends on the opinions on the trusted peers in their network. (Kempe et al., 2003) have followed up on the challenge of maximizing viral information spread by evaluating several algorithms given various models of adoption we discuss next.

3 Completed work: Network structure and evolution

The first part of the thesis presents properties of time evolving networks we discovered. This motivated the development of new probabilistic generative models and algorithms to fit them to real networks.

3.1 Properties of evolving networks

We studied a range of different networks, from several domains, and focused specifically on the way in which fundamental structural properties of networks vary with time. Our results suggest that two fundamental and commonly made assumptions about network evolution need to be reassessed:

(A) *Constant average degree assumption*: The average node degree in the network remains constant over time. (Or equivalently, the number of edges grows linearly in the number of nodes.) (Albert and Barabasi, 1999, Newman, 2003)

(B) *Slowly growing diameter assumption*: The diameter is a slowly growing function of the network size, as in "small world" graphs. (More precisely, the diameter of the graph increases logarithmically in the number of nodes in the graph.) (Albert et al., 1999, Broder et al., 2000, Milgram, 1967, Watts and Strogatz, 1998)

3.1.1 Densification Power Law

In contrast to conventional wisdom we found that networks from various domains *densify* over time with the number of edges growing super-linearly in the number of nodes. This means that the later the node joins the network the more edges it will create. Furthermore, the network is not arbitrarily densifying but it follows a *Densification Power Law* – the growing network maintains the power-law relationship between the number of nodes and the number of edges over time:

$$e(t) \propto n(t)^a,$$

where $e(t)$ and $n(t)$ are the number of edges and nodes of the graph at time t, and a is a *Densification exponent* that lies strictly between 1 and 2. Exponent $a = 1$ corresponds to constant average degree over time (which was assumed so far), while $a = 2$ corresponds to an extremely dense graph.

For example, figure 1 shows the Densification Power Law for a large physics citation network, which was obtained from `arXiv.org`. The network has $29,555$ nodes and $347,268$ edges and spans a period of 10 years. A second dataset is the Autonomous Systems (AS), which can be thought of as a graph of the internet. We have 735 daily instances for a period of over 2 years, and the largest instance has $6,474$ nodes and $26,467$ edges. Notice the nontrivial densification exponents of $a = 1.7$ and $a = 1.2$. We refer the reader to (Leskovec et al., 2007c) for more examples of densifying networks.

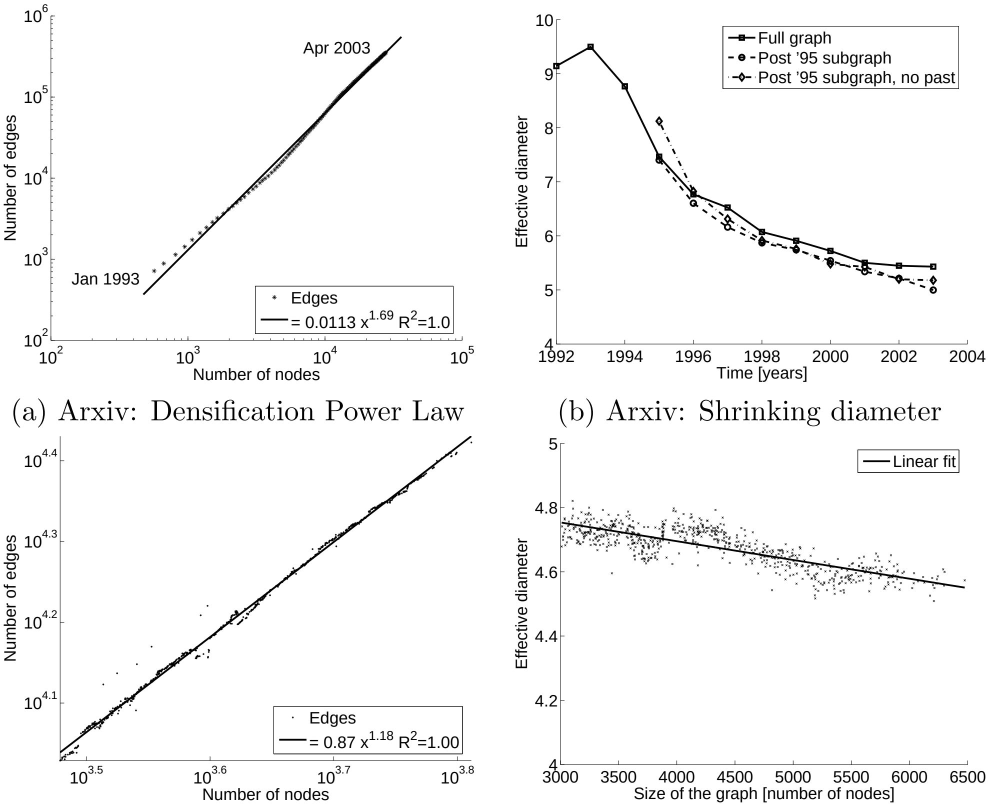
3.1.2 Shrinking diameters

A second, even more surprising observation is that the average distance between nodes in a graph *shrinks* over time rather than increases slowly as a function of the number of nodes, as it is commonly believed. This result is particularly surprising since it moves the long-running debate over exactly how slowly the graph diameter *grows* (Bollobas and Riordan, 2004, Chung and Lu, 2002), to the need to revisit standard models so as to produce graphs in which the effective diameter is *shrinking* over time.

Figure 1 shows example of Shrinking Diameters for a large physics citation network and the Autonomous Systems (AS). Notice the gradual decrease in effective diameter as the network grows. Again, more examples of networks with shrinking diameters can be found in (Leskovec et al., 2007c).

3.1.3 Densification and degree distribution

As we saw many networks give rise to heavy tailed (power law) degree distribution. Next, we present analysis of the relation between the densification and the power-law degree distribution over time. We show they are fundamentally related, and that there are two regimes where densification occurs: (a) power-law degree distribution evolves over time to



(a) Arxiv: Densification Power Law

(b) Arxiv: Shrinking diameter

(c) AS: Densification Power Law

(d) AS: Shrinking diameter

Figure 1: (a) Densification Power Law: number of edges $e(t)$ versus number of nodes $n(t)$, in log-log scales. Notice non-trivial Densification exponents $a = 1.69$ and $a = 1.2$. (b) The "effective" diameter over time. Notice it is shrinking as the graph grows.

allow for densification. (b) power-law degree exponent remains constant over time. In this case the Densification Power Law is the consequence of the fact that a power-law distribution with exponent $\gamma < 2$ has no finite expectation (Newman, 2005), and thus the average degree grows as the degree exponent remains constant.

We formalize these with the following theorems (Leskovec et al., 2007c):

Theorem 3.1 (Leskovec et al. (2007c)) *Given a time evolving graph on n nodes that evolves according to Densification Power Law with exponent $a > 1$ and has a Power-Law degree distribution with exponent $\gamma_n > 2$, then the degree exponent γ_n evolves with the number of nodes n as*

$$\gamma_n = \frac{4n^{a-1} - 1}{2n^{a-1} - 1}$$

Theorem 3.2 (Leskovec et al. (2007c)) *In a temporally evolving graph with a power-law degree distribution having constant degree exponent* γ *over time, the Densification Power Law exponent* a *is:*

$$\begin{aligned} a &= 1 && \text{if } \gamma > 2 \\ &= 2/\gamma && \text{if } 1 \le \gamma \le 2 \\ &= 2 && \text{if } \gamma < 1 \end{aligned}$$

We also found cases of real world networks that follow the results of the above theorems. We find that citation networks densify by flattening (decreasing) degree exponent (Theorem 3.1), and that the Email networks densify by having constant degree exponent, $\gamma = 1.8 < 2$ (Theorem 3.2). Details on the analysis and experiments can be found in (Leskovec et al., 2007c, Section 5).

3.2 Explanatory models

What underlying process causes a graph to systematically densify and experience a decrease in effective diameter even as its size increases? Existing graph generation models (Albert and Barabasi, 1999, Newman, 2003) do not exhibit these types of behavior. This question motivates the next part of our work: we developed two families of probabilistic generative models for graphs that do capture these properties.

3.2.1 Community guided attachment

The first model, which we refer to as *Community Guided Attachment* (Leskovec et al., 2005b), shows that a decomposition of the nodes into a nested set of communities, such that the difficulty of forming links between communities increases with the distance in the hierarchy, naturally explains the *Densification Power Law* with any desired exponent. In short, self-similarity itself leads to the Densification Power Law. The proofs, further details and extension of the model can be found in our papers (Leskovec et al., 2007c).

We represent the recursive structure of communities-within-communities as a tree Γ, of height H. We show that even a simple, perfectly balanced tree of constant fanout b is enough to lead to a densification power law, and so we will focus the analysis on this basic model.

The nodes V in the graph we construct will be the leaves of the tree; that is, $n = |V|$. (Note that $n = b^H$.) Let $h(v, w)$ define the standard tree distance of two leaf nodes v and w: that is, $h(v, w)$ is the height of their least common ancestor.

We construct a random graph on a set of nodes V by specifying the probability that v and w form an edge as a function f of $h(v, w)$. We refer to this function f as the *Difficulty Function.* What should be the form of f? Clearly, it should decrease with h; but there are many forms such a decrease could take.

The form of f that works best for our purposes comes from the self-similarity arguments: We would like f to be scale-free; that is, $f(h)/f(h-1)$ should be level-independent and thus constant. The only way to achieve level-independence is to define $f(h) = f(0)c^{-h}$. Setting $f(0)$ to 1 for simplicity, we have:

$$f(h) = c^{-h} \tag{1}$$

where $c \geq 1$. We refer to the constant c as the *Difficulty Constant*. Intuitively, cross-communities links become harder to form as c increases.

This completes our development of the model, which we refer to as *Community Guided Attachment*: If the nodes of a graph belong to communities-within-communities, and if the cost for cross-community edges is scale-free (Eq. (1)), the Densification Power Law follows naturally. No central control or exogenous regulations are needed to force the resulting graph to obey this property. In short, self-similarity itself leads to the Densification Power Law:

Theorem 3.3 (Leskovec et al. (2005b)) *In the Community Guided Attachment random graph model just defined, the expected average out-degree $\bar{d}$ of a node is proportional to:*

$$\begin{aligned} \bar{d} &= n^{1-\log_b(c)} && \textit{if } \ 1 \leq c \leq b \\ &= \log_b(n) && \textit{if } \ c = b \\ &= \textit{constant} && \textit{if } \ c > b \end{aligned}$$

The proof and further extensions of the basic model can be found in (Leskovec et al., 2005b, Theorem 1).

3.2.2 Forest Fire model

Community Guided Attachment and its extensions show how densification can arise naturally, and even in conjunction with heavy-tailed in-degree distributions. However, it is not a rich enough class of models to capture all the properties in our network datasets. In particular, we would like to capture both the shrinking effective diameters that we have observed, as well as the fact that real networks tend to have heavy-tailed out-degree distributions (though generally not as skewed as their in-degree distributions). The Community Guided Attachment models do not exhibit either of these properties.

Specifically, our goal is as follows. Given a (possibly empty) initial graph G, and a sequence of new nodes $v_1 \ldots v_k$, we want to design a simple randomized process to successively link v_i to nodes of G ($i = 1, \ldots k$) so that the resulting graph G_{final} will obey all of the following patterns: heavy-tailed distributions for in- and out-degrees, the Densification Power Law, and shrinking effective diameter.

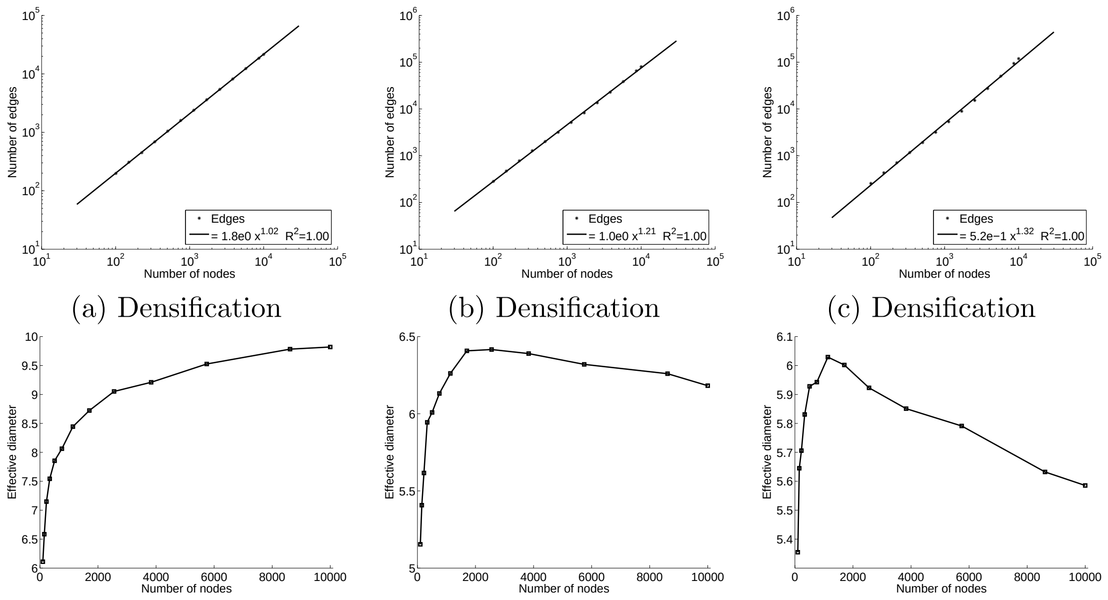
We introduce the *Forest Fire Model* (Leskovec et al., 2005b), which is capable of producing all these properties. To set up this model, we begin with some intuition that also underpinned Community Guided Attachment: nodes arrive over time; each node has a "center of gravity" in some part of the network; and its probability of linking to other nodes decreases rapidly with their distance from this center of gravity. However, we add to this picture the notion that, occasionally, a new node will produce a very large number of out-links. Such nodes will help cause a more skewed out-degree distribution; they will also serve as "bridges" that connect formerly disparate parts of the network, bringing the diameter down.

Following this plan, we now define the most basic version of the model. Essentially, nodes arrive one at a time and form out-links to some subset of the earlier nodes; to form out-links, a new node v attaches to an *ambassador* node w in the existing graph, and then begins "burning" links outward from w, linking with a certain probability to any new node it discovers. One can view such a process as intuitively corresponding to a model by which an author of a paper identifies references to include in the bibliography. He or she finds a first paper to cite, chases a subset of the references in this paper (modeled here as random), and continues recursively with the papers discovered in this way. Depending on the bibliographic aids being used in this process, it may also be possible to chase back-links to papers that cite the paper under consideration.

Despite the fact that there is no explicit hierarchy in the Forest Fire Model, as there was in Community Guided Attachment, there are some subtle similarities between the models. Where a node in Community Guided Attachment was the child of a parent in the hierarchy, a node v in the Forest Fire Model also has an "entry point" via its chosen ambassador node w. Moreover, just as the probability of linking to a node in Community Guided Attachment decreased exponentially in the tree distance, the probability that a new node v burns k successive links so as to reach a node u lying k steps away is exponentially small in k.

In fact, our Forest Fire Model combines the flavors of several older models, and produces graphs qualitatively matching their properties. We establish this by simulation, as we describe below, but it is also useful to provide some intuition for why these properties arise.

- *Heavy-tailed in-degrees.* Our model has a "rich get richer" flavor: highly linked nodes can easily be reached by a newcomer, no matter which ambassador it starts from.
- *Communities.* The model also has a "copying" flavor: a newcomer copies several of the neighbors of his/her ambassador (and then continues this recursively).
- *Heavy-tailed out-degrees.* The recursive nature of link formation provides a reasonable chance for a new node to burn many edges, and thus produce a large out-degree.
- *Densification Power Law.* A newcomer will have a lot of links near the community of his/her ambassador; a few links beyond this, and significantly fewer farther away. Intuitively, this is analogous to the Community Guided Attachment, although without an explicit set of communities.
- *Shrinking diameter.* It is not a priori clear why the Forest Fire Model should exhibit a shrinking diameter as it grows. Graph densification is helpful in reducing the diameter, but it is important to note that densification is certainly not enough on its own to imply shrinking diameter. For example, the Community Guided Attachment model obeys the Densification Power Law, but our experiments also show that the diameter slowly increases (not shown here for brevity).



(a) Densification (b) Densification (c) Densification

(a) Diameter (b) Diameter (c) Diameter

Figure 2: The Densification Power Law plot and the diameter for Forest Fire model. Column 1: sparse graph ($a = 1.01 < 2$), with increasing diameter. Column 2: (most realistic case:) densifying graph ($a = 1.21 < 2$) with slowly decreasing diameter. Column 3: densifying graph ($a = 1.32 < 2$) with decreasing diameter.

Figure 2 shows the evolution of the network for different values of parameters. Notice Forest Fire model produces graphs of various densifications and levels of shrinking diameter, while also generating networks with power-law degree distributions (plots not shown for brevity, see (Leskovec et al., 2007c)).

3.3 Kronecker graphs generative model

Our next goal is to develop an analytically tractable model of network generation and evolution which can easily be analyzed and fitted to real networks.

Next, more sophisticated model, exhibits the *full* range of properties. It is based on a non-standard matrix operation, the *Kronecker product*. Intuitively, communities in the graph grow recursively, with nodes recursively getting expanded into miniature copies of the community. Nodes in the subcommunity then link among themselves and to the nodes in different communities.

The beauty of *Kronecker Graphs* (Leskovec et al., 2005a) is that they are mathematically very tractable. We can prove that they obey *all static* and *dynamic* patterns that were observed in large real-world networks: heavy-tailed distributions for in-degree, out-degree, eigenvalues and eigenvectors, constant/shrinking diameter and densification power law.

First, we introduce deterministic version of Kronecker Graphs which we will later extend to Stochastic Kronecker Graphs for which we also developed scalable algorithm for parameter estimation.

3.3.1 Deterministic Kronecker Graphs

The main idea is to create self-similar graphs, recursively. We begin with an *initiator* graph G_1, with N_1 nodes, and by recursion we produce successively larger graphs $G_2 \dots G_n$ such that the k^{th} graph G_k is on $N_k = N_1^k$ nodes.

If we want these graphs to exhibit a version of the Densification Power Law, then G_k should have $E_k = E_1^k$ edges. This is a property that requires some care in order to get right, as standard recursive constructions (for example, the traditional Cartesian product or the construction of (Barabasi et al., 2001)) do not satisfy it.

As it turns out the *Kronecker product* is a perfect tool for this goal. It is defined as:

Definition 1 (Kronecker product of matrices) *Given two matrices* $\mathbf{U} = [u_{i,j}]$ *and* $\mathbf{V}$ *of sizes* $n \times m$ *and* $n' \times m'$ *respectively, the Kronecker product matrix* $\mathbf{S}$ *of dimensions* $(n * n') \times (m * m')$ *is given by*

$$\mathbf{S} = \mathbf{U} \otimes \mathbf{V} \doteq \begin{pmatrix} u_{1,1}\mathbf{V} & u_{1,2}\mathbf{V} & \dots & u_{1,m}\mathbf{V} \\ u_{2,1}\mathbf{V} & u_{2,2}\mathbf{V} & \dots & u_{2,m}\mathbf{V} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n,1}\mathbf{V} & u_{n,2}\mathbf{V} & \dots & u_{n,m}\mathbf{V} \end{pmatrix} \tag{2}$$

Kronecker product of two graphs is defined as Kronecker product of their adjacency matrices:

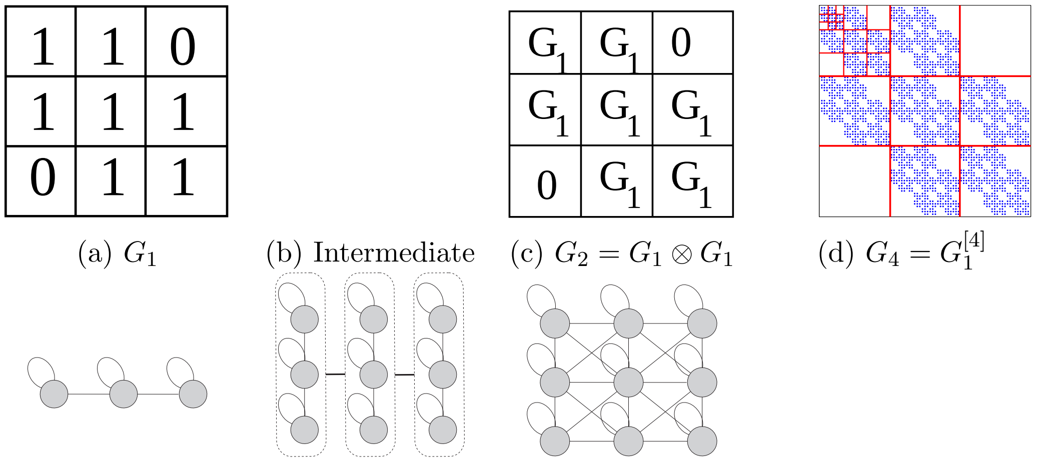
Definition 2 (Kronecker product of graphs) *Let* G *and* H *be graphs with adjacency matrices* $A(G)$ *and* $A(H)$ *respectively, then the Kronecker product* $G \otimes H$ *is defined as the graph with adjacency matrix* $A(G) \otimes A(H)$*.*

And, we denote k^{th} Kronecker power of G_1 as $G_1^{[k]}$ (abbreviated to G_k), where $G_k = G_1^{[k]} = G_{k-1} \otimes G_1$:

Definition 3 (Kronecker power) *The* k^{th} *power of* G_1 *is defined as the matrix* $G_1^{[k]}$ *(abbreviated to* G_k*) , such that:*

$$G_1^{[k]} = G_k = \underbrace{G_1 \otimes G_1 \otimes \dots G_1}_{k\ times} = G_{k-1} \otimes G_1$$

Figure 3 shows the recursive construction of Kronecker graphs. We start with G_1, a 3-node chain, and Kronecker power it to obtain G_2. The self-similar nature of the Kronecker graphs is clear: To produce G_k from G_{k-1}, we "expand" (replace) nodes of G_{k-1} by copies of G_1, and join the copies according to the adjacencies in G_{k-1} (see fig. 3). One can imagine this by positing that communities in the graph grow recursively, with nodes in the community recursively getting expanded into miniature copies of the community. Nodes in the sub-community then link among themselves and to nodes from other communities.



(a) G_1 (b) Intermediate (c) $G_2 = G_1 \otimes G_1$ (d) $G_4 = G_1^{[4]}$

Figure 3: *Kronecker multiplication:* Top row: structure of adjacency matrices. Bottom: corresponding graphs – "3-chain" and its Kronecker product with itself; each of the nodes gets expanded into 3 nodes, which are then linked.

3.3.2 Stochastic Kronecker Graphs

We also define a stochastic version of Kronecker Graphs. The difference is that now the initiator matrix is stochastic: we start with a $N_1 \times N_1$ *probability matrix* $\Theta = [\theta_{ij}]$, where the element $\theta_{ij} \in [0, 1]$ is the probability that edge (i, j) is present. We compute the k^{th} Kronecker power $\mathcal{P} = \Theta^{[k]}$; And then for each $p_{uv} \in \mathcal{P}$, include edge (u, v) with probability p_{uv}.

Stochastic Kronecker Graphs are thus parameterized by the $N_1 \times N_1$ *probability (parameter) matrix* Θ. The probability p_{uv} of an edge (u, v) occurring in k-th Kronecker power $\mathcal{P} = \Theta^{[k]}$.

To sample a Kronecker graph G, i.e. obtain a realization, from $\mathcal{P}$ we perform the following procedure: for each $[p_{ij}] \in \mathcal{P}$ we include an edge (i, j) in G with probability p_{ij}, *i.e.* we have a Bernoulli edge generation model.

3.3.3 Properties of Kronecker Graphs

Kronecker graphs have a rich set of properties that are also found in real networks. More specifically we show that Kronecker graphs have the following properties Leskovec et al. (2005b):

Theorem 3.4 (Multinomial degree distribution) *Kronecker graphs have multinomial degree distributions, for both in- and out-degrees.*

Note that multinomial distribution with a proper choice of parameters can be made to behave as heavy-tailed (power-law) distribution. For example, see Figure 4(a) and also (Leskovec et al., 2005a).

Theorem 3.5 (Multinomial eigenvalue distribution) *The Kronecker graph G_k has a multinomial distribution for its eigenvalues.*

Theorem 3.6 (Multinomial eigenvector distribution) *The components of each eigenvector of the Kronecker graph G_k follow a multinomial distribution.*

Theorem 3.7 (Densification) *Kronecker graphs follow the Densification Power Law (DPL) with densification exponent $a = \log(E_1)/\log(N_1)$.*

Theorem 3.8 (Diameter) *If G and H each have diameter at most d, and each has a self-loop on every node, then the Kronecker product $G \otimes H$ also has diameter at most d.*

Further details on theorems and proofs can be found in (Leskovec et al., 2005a).

As we will see in next section Kronecker graphs can also be fit to real data and they seem to be a model of just the right complexity, *i.e.* not too big parameter space while still maintaining rich expressive power, to capture properties of real graphs.

3.3.4 Estimating parameters of Kronecker graphs

Kronecker graphs are promising, since they obey many patterns found in real life networks and have very intuitive and informative parameters — the whole model is captured by the "initiator" (or "seed") graph. Given a set of constraints (patterns we want to match) we are searching for the initiator graph. Our goal is to compute the likelihood over a set of possible initiator graphs and seek the most likely one.

Stochastic graph models introduce probability distributions over graphs. A generative model assigns probability $P(G)$ to every graph G. $P(G)$ is the *likelihood* that a given model generated graph G. We concentrate on Stochastic Kronecker Graph model, and consider fitting it to a real graph G. We use a maximum likelihood approach, *i.e.* we aim to find parameter values Θ that maximize the $P(G)$ under the model. This presents several challenges:

- **Model selection:** A graph is a single structure, and not a set of items drawn i.i.d. from some distribution. So one can not split it into independent training and test sets. The fitted parameters will thus be best to generate a *particular* instance of a graph. Also, overfitting is an issue since a more complex model usually fits better.

- **Node labeling:** The second issue is the node ordering or node labeling. Graph G has a set of N nodes, and each node has unique index (label, number). Labels do not carry any particular meaning. One can think of this as a graph is first generated and then the labels are randomly assigned to the nodes. This means that two isomorphic graphs that have different node labeling should have the same likelihood. So to compute the likelihood one has to consider all node labelings $P(G) = \sum_\sigma P(G|\sigma)P(\sigma)$, where the sum is over all permutations σ of N nodes.

- **Likelihood estimation:** Calculating $P(G|\sigma)$ naively takes $O(N^2)$ by simply evaluating the probability of each edge in the graph adjacency matrix. The challenge is averaging over the *super-exponentially* many permutations which is computationally intractable, and thus one has to reside to simulation and sampling. For real graphs even calculating $P(G|\sigma)$ in $O(N^2)$ is infeasible.

As the problem is introduced there are several difficulties. First, we assume gradient descent type optimization will work, *i.e.* the problem does not have (too many) local minima. Second, we are summing over exponentially many permutations, i.e. node labelings. Third, the evaluation of the $P(G|\sigma)$ takes $O(N^2)$, and needs to be evaluated $N!$ times.

Observation 1 *Naively calculating the likelihood $P(G|\Theta)$ of a Stochastic Kronecker Graph with parameters Θ takes $O(N!N^2)$, where N is the number of nodes in G.*

We developed KronFit (Leskovec and Faloutsos, 2007), an algorithm for estimating parameters Θ given a real graph G that runs in *linear* time.

Observation 2 *Given a graph G,* KronFit *estimates the parameters Θ of Stochastic Kronecker Graph in time $O(E)$, where E is the number of edges in G.*

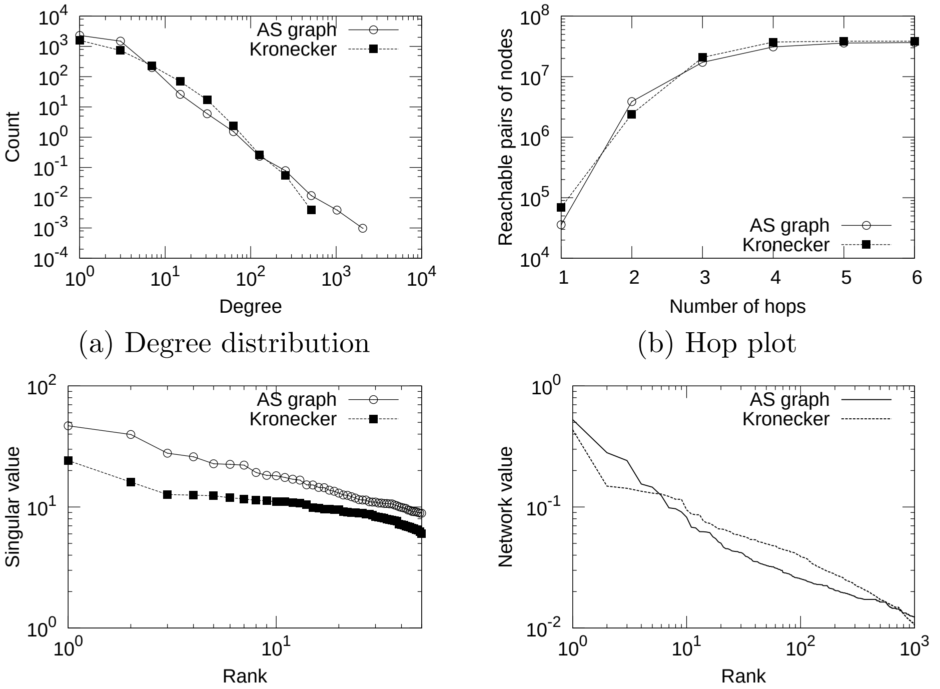
We use simulation techniques to avoid the super-exponential sum over the node labelings. By exploiting the structure of Kronecker matrix multiplication we can evaluate $P(G|\sigma)$ in *linear* time $O(E)$. And since real graphs are *sparse*, *i.e.* the number of edges is of the same order as the number of nodes, this makes the fitting of the Kronecker model to large graphs tractable (Leskovec and Faloutsos, 2007).

3.3.5 Experiments with Kronecker Graphs

Next, we present a series of experiments that show that KronFit is able to recover true parameters when given a synthetic graph, and that synthetic graphs generated from the estimated parameters fit the real graphs well.

Optimization space: In Kronecker graphs permutations of the parameter matrix Θ all have the same likelihood. This means that the maximum likelihood optimization problem is not convex, but rather has several global minima. To check for the presence of other local minima where gradient descent could get stuck we run the following experiment: we generated 100 synthetic Kronecker graphs on 16,384 (2^{14}) nodes and 1.4 million edges on average, with a randomly chosen 2×2 parameter matrix Θ^*. For each of the 100 graphs we start gradient descent from a different random location Θ', and try to recover Θ^*. In **98%** of the cases the descent converged to the true parameters. Many times the algorithm converged to a different global minima, *i.e.* permuted true parameter values. This suggests surprisingly nice structure of the optimization problem: it seems it behaves like a convex optimization problem with many equivalent global minima.

Fitting to real-world graphs: We also present experiments of fitting the Kronecker Graphs model to real-world graphs. Given a real graph G we aim in discovering most likely parameters $\hat{\Theta}$ that ideally would generate a synthetic graph K having same properties as G. This assumes that Kronecker Graphs is a good model for real graphs, and that KronFit is able to recover good parameters. We take a real graph G, find parameters $\hat{\Theta}$ using



(a) Degree distribution

(b) Hop plot

(c) Scree plot

(d) "Network" value

Figure 4: *Autonomous Systems:* Overlayed patterns of real graph and the fitted Kronecker graph. Notice that the fitted Kronecker graph matches patterns of the real graph.

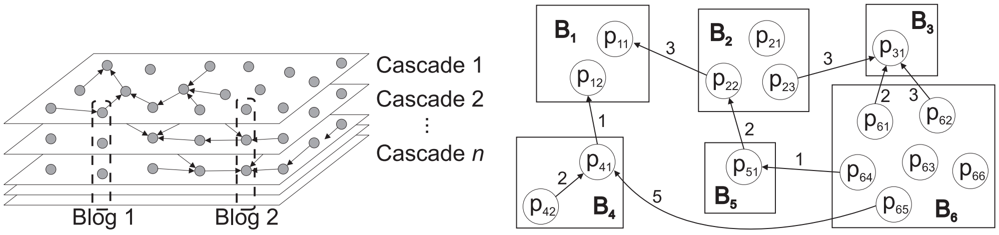
KronFit, generate a synthetic graph K using $\hat{\Theta}$, and compare their properties that we introduced in section 2.

Figure 4 shows properties of Autonomous Systems graph (6,474 nodes, 26,467 edges), and compares them with the properties of a synthetic Kronecker graph generated using the fitted parameters $\hat{\Theta}$ of size 2×2. Notice that properties of both graphs match really well.

This is a nice result since it also shows that through the optimization of the maximum likelihood the graphs also match in several other properties even though we are not directly optimizing over them.

Autonomous Systems network is undirected, and the fitted parameter matrix $\hat{\Theta} = [.98, .58; .58, .06]$ reveals this. This means that without a priori biasing the fitting towards undirected graphs, the recovered parameters obey this. Fitting AS graph from a random set of parameters took less than 20 minutes on a standard desktop PC. This is a significant speedup over (Bezáková et al., 2006), where by using a similar permutation sampling approach for calculating the likelihood of a preferential attachment model on similar AS graph took about two days on a cluster of 50 machines.

In contrast to earlier work, our work has the following novelties: (a) it is among the few that estimates the parameters of the chosen generator (b) it is among the few that has a concrete measure of goodness of the fit (namely, likelihood) (c) it avoids the quadratic complexity of computing the likelihood by exploiting the properties of the "Kronecker graphs"



(a) Cascades as layers

(b) Cascades as graphs

Figure 5: Two views on the formation of information cascades on the blogosphere.

(d) it avoids the factorial explosion of the correspondence problem, by using Metropolis sampling.

The benefits of fitting a Kronecker graph model into a real graph are several: *Extrapolation*: Once we have the Kronecker generator Θ for a given real matrix G (such that G is mimicked by $\Theta^{[k]}$), a larger version of G would be generated by $\Theta^{[k+1]}$. *Sampling*: Similarly, if we want a realistic sample of the real graph, we could use a smaller exponent in the Kronecker exponentiation, like $\Theta^{[k-1]}$. *Anonymization*: Since $\Theta^{[k]}$ mimics G, we can publish $\Theta^{[k]}$, without revealing information about the nodes of the real graph G.

4 Completed work: Network cascades

The second part of the thesis focuses on the notion of *information cascades* — a phenomena where an action or idea becomes widely adopted due to influence by others, as opposed to individual reasoning in isolation. We formally define a cascade as a graph where the nodes are agents and a directed edge (i, j, t) indicates that a node i influenced a node j at time t.

We consider three examples of cascade formation and propagation in networks:

- First, we present results on cascades in a large viral marketing network, where people recommend products to each other and we study the spread and success of recommendations over the network.
- Second, we consider the tracking of a large population of blogs over a long period of time and observe the propagation of information between the blogs.
- Third, we study the propagation of infectious water in large real water distribution networks, and ask the question of where to place a limited number of sensors so the disease outbreaks will be detected early.

Blogs (weblogs) are web sites that are updated on a regular basis. Often times individuals use them for online diaries and social networking; other times news sites have blogs for timely stories. Blogs are composed of time-stamped posts, and posts typically link each other, as well as other resources on the Web.

For example, figure 5 shows two alternative views of information cascades that may occur on the blogosphere. In figure 5(a) each circle represents a blog post, and all circles at the same vertical position belong to the same blog. Often blog posts refer to each other using hyper-links. Given that the posts are time-stamped and usually not updated, we can trace their linking patterns all the way to the source. It is easy to identify the flow if information from the source post to the followers and followers of the followers. So, each layer represents a different information cascade (information propagation graph). Figure 5(b) gives an alternative view. Here posts (represented as circles) inside a rectangle belong to the same blog. Similarly, the information cascades correspond to connected components of the posts in the graph, *e.g.* posts p_{12}, p_{41}, p_{42} and p_{65} all form a cascade, where p_{12} is the *cascade initiator*.

Observing such behavior on the blogosphere or in the viral marketing poses several interesting questions: What kinds of cascades arise frequently in real life? Are they like trees, stars, or something else? And how do they reflect properties of their underlying network environment? How fast does the information spread? Do certain nodes have specific propagation patterns? What are the most important nodes to target if we want to spread the information over the network?

In addition to observing rich cascades and propagation (Leskovec et al., 2006b) we go a step further and analyze the effectiveness and dynamics of product recommendations in causing purchases (Leskovec et al., 2006a, 2007a). To our knowledge this was the first study to directly observe the effectiveness of person to person word of mouth advertising for hundreds of thousands of products. Similarly, for blogs we (Leskovec et al., 2007e) are the first to perform a large study of cascading behavior in large blog networks.

4.1 Cascades in viral marketing

We study a recommendation network consisting of 4 million people who made 16 million recommendations on half a million products from a large on-line retailer. Each time a person purchases a book, music, DVD, or video tape she is given the option to send an email recommending the item to her friends. The first recipient to purchase the item receives a discount and the sender of the recommendation receives a referral credit.

Figure 6 shows two typical product recommendation networks. Most product recommendation networks consist of a large number of small disconnected components where we do not observe cascades. Then there is usually a small number of relatively small components where we observe recommendations propagating. We also notice bursts of recommendations and collisions (figure 6(b)). Some individuals send recommendations to many friends which results in star-like patterns in the graph.

4.1.1 Cascading patterns

We consider the problem of finding patterns of recommendations in a large social network. We ask the following questions: How does the influence propagate? What does it look like?

In order to analyze the data, we developed new methods and algorithms. First, we identify cascades, *i.e.* graphs where incoming recommendations influenced purchases and further recommendations. Next, we enumerate and count the cascade subgraphs. Graph isomorphism and enumeration are both computationally very expensive, so we developed new algorithms for approximate graph isomorphism resolution (Leskovec et al., 2006b).

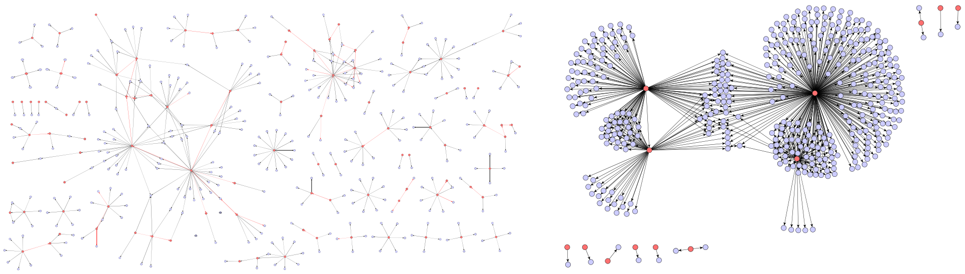


Figure 6: Examples of two product recommendation networks. Left: First aid study guide. Notice many small disconnected cascades. Right: Japanese graphic novel (manga). Notice a large, tight community.

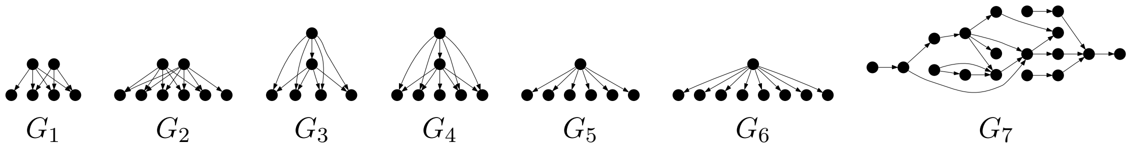


Figure 7: Typical classes of cascades. G_1, G_2: nodes recommending to the same set of people, but not each other. G_3, G_4: nodes recommending to same community. G_5, G_6: a flat cascade. G_7: a large propagation of recommendations.

In our multi-level approach the computational complexity (and accuracy) of the graph isomorphism resolution depends on the size of the graph. This property makes the algorithm scale nicely to large datasets.

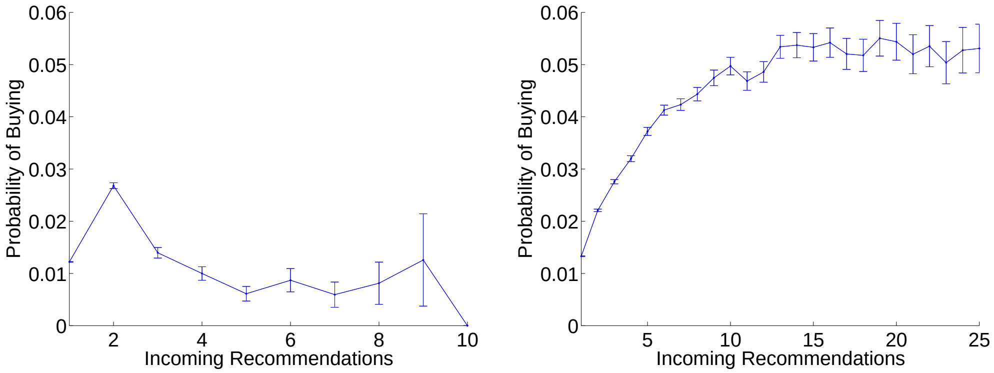
We found that the distribution of sizes and depths of cascades follows a power law. Generally, cascades tend to be shallow, but occasional large bursts can occur. Cascades are mainly tree-like, but we observe variability in connectivity and branching across different products groups. Figure 7 shows some typical examples of how the influence propagates over the recommendation network.

In addition to observing rich cascades and propagation we go a step further and analyze the effectiveness and dynamics of product recommendations in causing purchases.

4.1.2 Implications for viral marketing

We established how the recommendation network grows over time and how effective it is from the viewpoint of the sender and receiver of the recommendations. We can see what kind of product is more likely to be bought as a result of recommendation, and describe the size of the cascade that results from recommendations and purchases. While on average recommendations are not very effective at inducing purchases and do not spread very far, there are product and pricing categories for which viral marketing seems to be very effective.

Figure 8 presents an example of our findings. We plot the probability of purchasing a product given the number of received recommendations. Surprisingly, as more book recommendations are received their success *decreases*. Success of DVD recommendations



(a) Book

(b) DVD

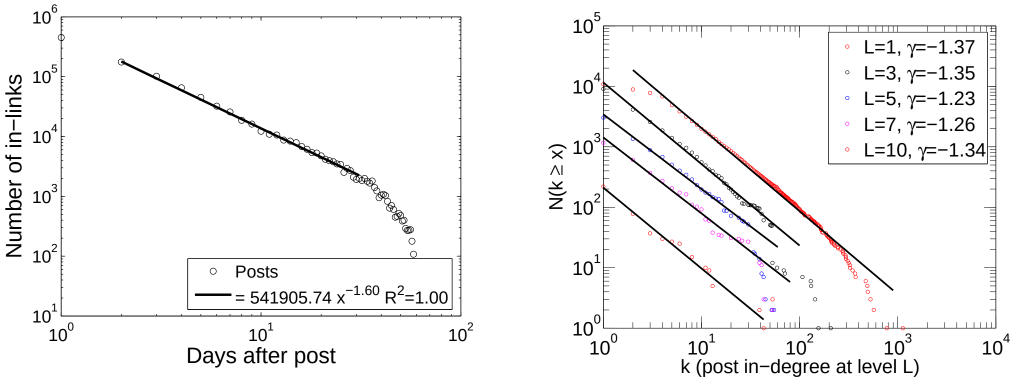
Figure 8: Probability of purchasing a product given the number of received recommendations. Notice the decrease in purchasing probability for books and saturation for DVDs.

saturates around 10 incoming recommendations. This means that after a person gets 10 recommendations they become immune to them – their probability of buying does not increase anymore. Traditional innovation diffusion models assume that an increasing number of infected contacts results in an increased likelihood of infection. Instead, we show that the probability of purchasing a product increases with the number of recommendations received, but then it quickly saturates. The result has important implications for viral marketing because providing too much incentive for people to recommend to one another can weaken the very social network links that the marketer is intending to exploit.

What determines the product's viral marketing success? We also developed a model which characterizes product categories for which recommendations are more likely to be accepted, and find that the numbers of nodes and receivers have negative coefficients, showing that successfully recommended products are actually more likely to be not so widely popular. It shows that more expensive and more recommended products have a higher success rate. These recommendations should occur between a small number of senders and receivers, which suggests a very dense recommendation network where lots of recommendations are exchanged between a small community of people. These insights could be of use to marketers — personal recommendations are most effective in small, densely connected communities enjoying expensive products. Refer to (Leskovec et al., 2007a) for more details.

4.2 Cascades on the blogosphere

Similarly to the viral marketing setting we analyzed cascades on the blogosphere. We address a set of related questions: What kinds of cascades arise frequently in real life? Are they like trees, stars, or something else? And how do they reflect properties of their underlying network environment?



Power-law popularity decay   In-degree at level L

Figure 9: Number of in-links vs. the days after the post in log-linear scale, after removing the day-of-the week effects. The power law fit has the exponent -1.5.

4.2.1 Shape of information cascades

We extracted our dataset from a larger set of blogs and posts from August and September 2005 (Glance et al., 2005). We were interested in blogs and posts that actively participate in discussions, so we biased our dataset towards the more active part of the blogosphere. We focused on the most-cited blogs and traced forward and backward conversation trees containing these blogs. This process produced a dataset of 2.5 million posts from $45,000$ blogs gathered over the three-month period. To analyze the data, we first create graphs of time-obeying propagation of links. Then, we enumerate and count all possible cascade subgraphs.

We find novel patterns, and the analysis of the results gives us insight into the cascade formation process. Most surprisingly, the popularity of posts drops with a *power law*, instead of exponentially, that one may have expected. We collect all in-links to a post and plot the number of links occurring after each day following the post. This creates a curve that indicates the rise and fall of popularity. Figure 9(a) shows number of in-links for each day following a post for all posts in the dataset The exponent of the power law is -1.5, which is exactly the value predicted by the model where the bursty nature of human behavior is a consequence of a decision based queuing process (Oliveira and Barabasi, 2005, Vazquez et al., 2006) – when individuals execute tasks based on some perceived priority, the timing of the tasks is heavy tailed, with most tasks being rapidly executed, whereas a few experience very long waiting times.

We also find that probability of observing a cascade on n nodes follows a Zipf distribution: $p(n) \propto n^{-2}$. Figure 9(b) plots the in-degree distribution of nodes at level L of the cascade. A node is at level L if it is L hops away from the root (cascade initiator) node. Notice that the in-degree exponent is stable and does not change much given the level in the cascade. This means that posts still attract attention (get linked) even if they are somewhat late in the cascade and appear towards the bottom of it.

We also found rich cascade patterns. Generally cascades are shallow but occasional

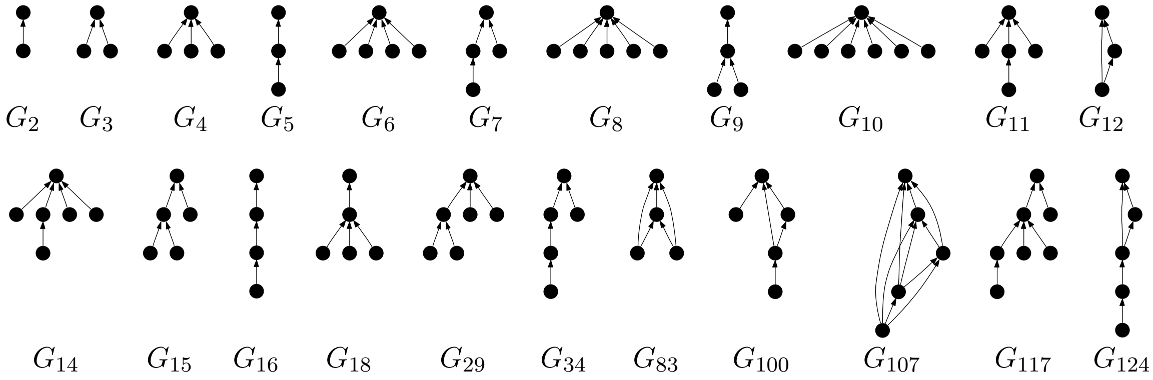


Figure 10: Common blog cascade shapes, ordered by the frequency of appearance.

large bursts also occur. The cascade sub-patterns shown on figure 10 reveal mostly small tree-like subgraphs; however we observe differences in connectivity, density, and the shape of cascades. Indeed, the frequency of different cascade subgraphs is not a simple consequence of differences in size or density; rather, we find instances where denser subgraphs are more frequent than sparser ones, in a manner suggestive of properties in the underlying social network and propagation process.

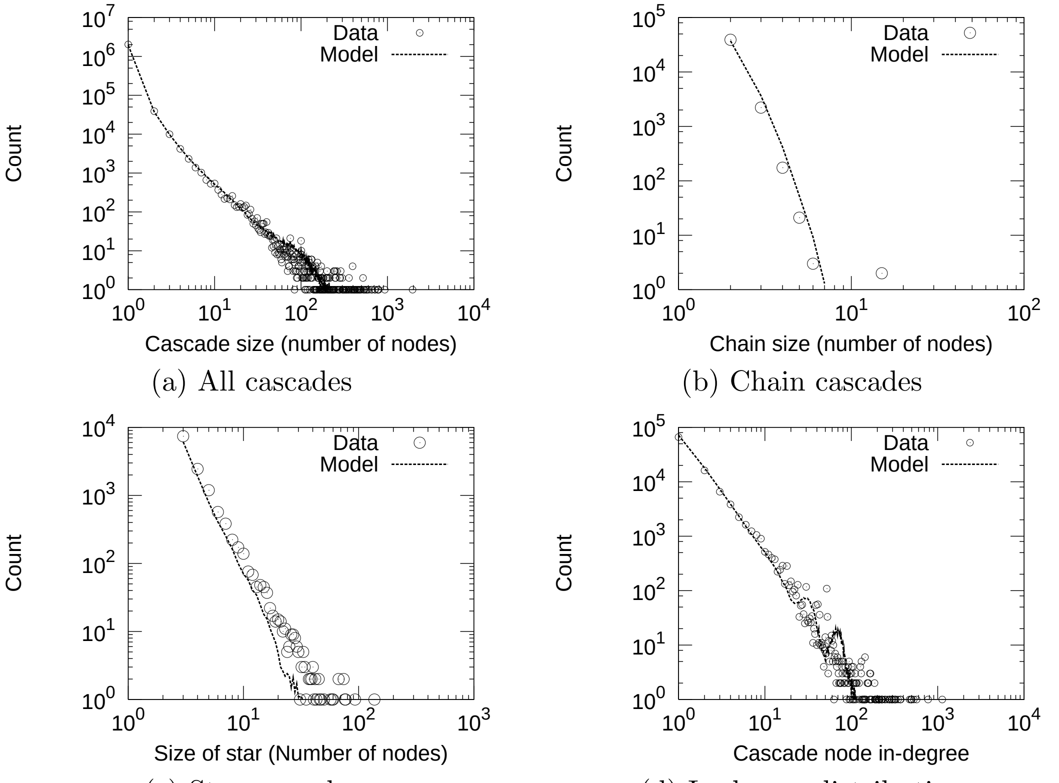
For example, we found that BoingBoing, which a very popular blog about amusing things, is engaged in many cascades. Actually, 85% of all BoingBoing posts were cascade initiators. The cascades generally did not spread very far but were wide (*e.g.*, G_{10} and G_{14} in Figure 10). On the other hand 53% of the posts from an influential political blog MichelleMalkin were cascade initiators, but the cascades here were deeper and generally larger (*e.g.*, G_{117} in Figure 10) than those of BoingBoing.

4.2.2 Model of information cascades

We also developed a conceptual model for generating information cascades that produces cascade graphs matching several properties of real cascades. Our model is intuitive and requires only a single parameter that corresponds to how interesting (easy spreading) the conversations in general on the blogosphere are.

Intuitively, cascades are generated by the following principle. A post is posted at some blog, other bloggers read the post, some create new posts, and link the source post. This process continues and creates a cascade. One can think of cascades as graphs created by the spread of a virus over the Blog network. This means that the initial post corresponds to infecting a blog. As the cascade unveils, the virus (information) spreads over the network and leaves a trail. To model this process we use a single parameter β that measures how infectiousness of the posts on the blogosphere. Our model is very similar to the SIS (susceptible – infected – susceptible) model from the epidemiology (Hethcote, 2000).

Figure 11 compares the cascades generated by the model with the ones found in the real blog network. Notice a very good agreement between the reality and simulated cascades in all plots. The distribution over cascade sizes is matched best. Chains and stars are slightly under-represented, especially in the tail of the distribution where the variance is



(a) All cascades

(b) Chain cascades

(c) Star cascades

(d) In-degree distribution

Figure 11: Comparison of the true data and the model. We plotted the distribution of the true cascades with circles and the estimate of our model with dashed line. Notice remarkable agreement between the data and the prediction of our simple model.

high. The in-degree distribution is also matched nicely, with an exception for a spike that can be attributed to a set of outlier blogs all with in-degree 52.

4.3 Node selection for early cascade detection

Next, we explore the general problem of detecting outbreaks in networks, where we are given a network and a dynamic process spreading over this network, and we want to select a set of nodes to detect the process as effectively as possible.

Many real-world problems can be modeled under this setting. Consider a city water distribution network, delivering water to households via pipes and junctions. Accidental or malicious intrusions can cause contaminants to spread over the network, and we want to select a few locations (pipe junctions) to install sensors, in order to detect these contaminations as quickly as possible.

Similarly with blogs we want to select a set of blogs to read (or retrieve) which are most up to date, *i.e.*, catch (link to) most of the stories that propagate over the blogosphere. Our goal is to select a small set of blogs (two in case of Figure 5) which "catch" as many cascades (stories) as possible. A naive, intuitive solution would be to select the big, well-known blogs. However, these usually have a large number of posts, and are time-consuming to read. We show, that, perhaps counter-intuitively, a more cost-effective solution can be obtained, by reading smaller, but higher quality, blogs, which our algorithm can find.

4.3.1 Node selection criteria

There are several possible criteria one may want to optimize in outbreak detection. For example, one criterion seeks to minimize *detection time* (*i.e.*, to know about a cascade as soon as possible, or avoid spreading of contaminated water). Similarly, another criterion seeks to minimize the *population affected* by an undetected outbreak (*i.e.*, the number of blogs referring to the story we just missed, or the population consuming the contamination we cannot detect). Optimizing these objective functions is NP-hard (Khuller et al., 1999), so for large, real-world problems, we cannot expect to find the optimal solution.

4.3.2 Exploiting submodularity

In our work (Leskovec et al., 2007d) we show that these and many other realistic outbreak detection objectives are *submodular* (Nemhauser et al., 1978), *i.e.*, they exhibit a diminishing returns property: Reading a blog (or placing a sensor) when we have only read a few blogs provides more new information, than reading it after we have read many blogs (placed many sensors). We find ways to exploit this submodularity property to *efficiently obtain* solutions which are *provably close* to the optimal solution. These guarantees are important in practice, since selecting nodes is expensive (reading blogs is time-consuming, sensors have high cost), and we desire solutions which are not too far from the optimal solution.

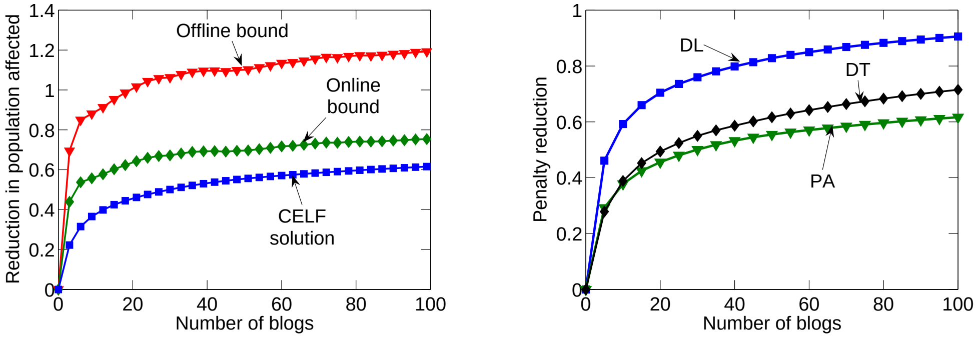
We also show that many objective functions for detecting outbreaks in networks are submodular, including detection time and population affected in the blogosphere and water distribution monitoring problems. We show that our approach also generalizes work by (Kempe et al., 2003) on selecting nodes maximizing influence in a social network.

We also exploit the submodularity of the objective (*e.g.*, detection time) to develop an efficient approximation algorithm, CELF, which achieves near-optimal placements (guaranteeing at least a constant fraction of the optimal solution), providing a novel theoretical result for non-constant node cost functions. CELF is up to 700 times faster than simple greedy algorithm. We also derive novel online bounds on the quality of the placements obtained by *any* algorithm.

4.3.3 Evaluation on water distribution and blog networks

We extensively evaluate our methodology on the applications introduced above – water quality and blogosphere monitoring. These are large real-world problems, involving a model of a water distribution network from the EPA with millions of contamination scenarios, and real blog data with millions of posts.

First, we evaluate the performance of CELF, and estimate how far from optimal the solution could be. Obtaining the optimal solution would require enumeration of $2^{45,000}$ subsets. Since this is impractical, we compare our algorithm to the bounds we developed.



(a) Solution quality

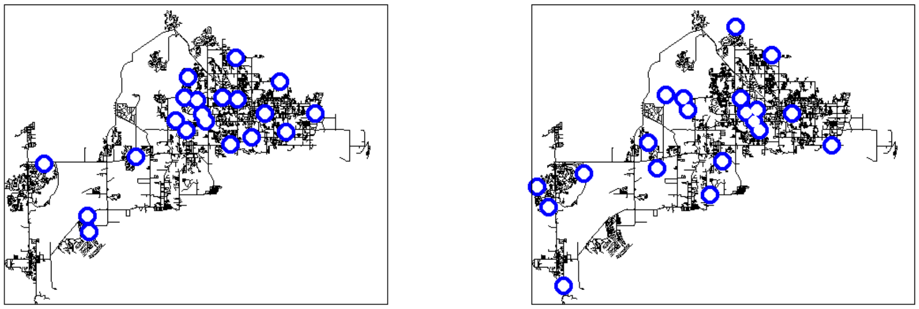
(b) Various Objective functions

Figure 12: Both plots show the solution quality vs. the number of selected sensors (blogs). (a) Performance of CELF algorithm and off-line and on-line bounds. Notice on-line bound is much tighter. (b) Compares different objective functions: detection likelihood (DL), detection time (DT) and population affected (PA).

Figure 12(a) shows scores for increasing budgets when optimized the Population affected criterion. As we select more blogs to read, the proportion of cascades we catch increases (bottom line). We also plot the two bounds. Notice the off-line bound (top line) is very loose. On the other hand, our on-line bound is much tighter than the traditional off-line bound.

In contrast to the off-line bound, our on-line bound is *algorithm independent*, and thus can be computed regardless of the algorithm used to obtain the solution. Since it is tighter, it gives a much better worst case estimate of the solution quality. For this particular experiment, we see that CELF works very well: after selecting 100 blogs, we are at most 13.8% away from the optimal solution. Similarly, figure 12(b) shows the performance using various objective functions. By using the on-line bound we also calculated that our results for all objective functions are at most 5% to 15% from optimal.

In August 2006, the Battle of Water Sensor Networks (BWSN) (Ostfeld et al., 2006) was organized as an international challenge to find the best sensor placements for a real metropolitan area water distribution network. In Figure 13 we show two 20 sensor placements obtained by our algorithm after optimizing Detection Likelihood and Population Affected, respectively. When optimizing the population affected, the placed sensors are concentrated in the dense high-population areas, since the goal is to detect outbreaks which affect the population the most. When optimizing the detection likelihood, the sensors are uniformly spread out over the network. Intuitively this makes sense, since according to BWSN challenge, outbreaks happen with same probability at every node. So, for Detection Likelihood, the placed sensors should be as close to all nodes as possible.



(a) Population Affected (b) Detection Likelihood

Figure 13: Water network sensor placements: (a) when optimizing Population Affected, sensors are concentrated in high population areas. (b) when optimizing Detection Likelihood, sensors are uniformly spread out.

5 Proposed work

We propose to extend the work we already completed, and also apply our methods to solve other problems related to graph mining.

The proposed work is composed of the following parts: (1) analysis and extension of Kronecker model to evolving networks; (2) empirical analysis and development of models for large online communication networks; (3) further study of information propagation and link creation in large online social networks. In parallel we also plan to publicly release a scalable graph mining library written in C++ that we developed during our research.

5.1 Research topic 1: Kronecker graphs

First, we propose to further work on various aspects of Kronecker Graphs. We propose to theoretically analyze their properties, and develop algorithms for fitting graphs over time. We also plan to extend model to be able to generate graphs with counts and any number of nodes.

- Theoretical analysis of Stochastic Kronecker graphs. In particular, we want to prove properties about the diameter of Stochastic Kronecker Graphs and their relation to Random Graphs of (Erdős and Rényi, 1960).

- Develop the models for fitting time evolving networks. We have ideas on developing a Hidden Markov type model, where the observable variable is a graph and the hidden variable corresponds to model parameters. We then allow the parameter matrix to slowly evolve over time. Besides for extrapolations to the future, the evolution of the parameter matrix will give us the means to interpret the evolution of the network.

- Extend Kronecker graphs model to generate networks with attributes on nodes and weights edges. The idea is to explore various generative processes that map the probability of an edge p_{ij} to the edge weight. So instead of having a set of Bernoulli trials that map from p_{ij} to actually observed edge, one could think ways for mapping p_{ij} to a weight, *e.g.* the number of messages or emails exchanged between a pair of nodes i, j.

- Extend Kronecker graphs to be able to generate graphs with any number of nodes. The idea here is to iteratively expand the graph with miniature copies of the initiator graph, *i.e.* instead of Kronecker powering the whole matrix at the same time which increases the number of nodes from N_1^k to N_1^{k+1}, one could for example pick a random element (i, j) of adjacency matrix and then Kronecker expand row i and column j of the current adjacency matrix. Each iteration of this process increases the number of nodes by N_1, and this way after k rounds we would have a graph on kN_1 nodes.

5.2 Research topic 2: Large online communication networks

Next, we propose to study static and temporal patterns in large communication networks, where the whole world communicates at once.

- We study large instant messenger communication networks with participants coming all over the world. These graphs have more than 200 million nodes and several billion edges (from 70GB to 3TB of data).
- Study how communication/network changes with the users demographics (gender, age, location, distance). We would use the communication as a global sensor. This data also gives us a perfect opportunity to measure the "six-degrees of separation" and other theories on a world scale. Results from these experiments then motivate the development of statistical models of communication and user demographics. These results could directly be applied to finding outliers (e.g., a scammer, pedophile, etc.)
- Also, the sheer scale of these data will lead to interesting technical and algorithmic questions on developing scalable algorithms for analyzing these huge networks.

5.3 Research topic 3: Nodes, links and information cascades

Last, we propose to further model and extend our analyses of information propagation in blog networks. In parallel with this we will also study link creation and adoption of a large professional social network, where we have available rich temporal information about the network from its start.

- With blogs the idea is to go beyond subgraph enumeration and identify real patterns by finding classes of graphs, *e.g.* near-trees, near-stars, etc. We expect to find cases where some blogs are "content providers", while others act as content "amplifiers" and make the content widely popular (*e.g.*, Slashdot). We aim to find characteristics and differences in linking patterns of content providers and amplifiers.
- The other interesting phenomena in social networks is propagation of trust. Here we have access to a large on-line social network of professional acquaintances. The

network is of moderate size with more than 7 million people and around 50 million edges between them. Besides the rich information we also have an opportunity of performing live experiments. We plan to study the link creation, *e.g.*, how do various user characteristics determine (geography, profession, structure of the network) the probability of a link? how do link invitations propagate over the network? Is there a critical mass of links when a person transitions from inviter to invitee?

5.4 Infrastructure: GraphGarden toolkit

We also plan to work on publicly releasing a general purpose graph mining and modeling library which was developed during our research. The library is written in C++ and it scales to massive graphs. The library contains more than 30 thousand lines of optimized code. Besides the library we will also create a set of accompanying utilities for analyzing properties of static and evolving networks, fitting models, calculating structural properties, analyzing cascades, etc.

5.5 Timeline

We plan to complete the proposed work according to the following tentative timeline:

- **May 2007:** Thesis proposal.
- **May 2007: Research topic 2** Research and modeling of Microsoft instant messenger communication network. Network has more than 200 million nodes and several million edges.
- **May – August 2007:** Research on on-line time evolving networks (summer internship at Yahoo Research).
- **September 2007: Infrastructure** Prepare the GraphGarden toolkit for public release along with the tutorial on graph mining that we submitted to ECML/PKDD conference.
- **October – December 2007: Research topic 1** Analyze properties of stochastic Kronecker graphs. Extend the Kronecker graphs model to fit time evolving networks, and relax the N_1^k nodes limitation.
- **January – May 2008: Research topic 3** Study of cascade formation and link prediction in the case of a large professional on-line social network.
- **May 2008:** Write the thesis.
- **June 2008:** Thesis defence.

6 Conclusion

The research focus of the proposed thesis is to analyze and model the structure, evolution and dynamics in large real-world networks. Our contributions so far are the following: We discovered novel properties of time evolving networks, namely Densification Power Law and Shrinking Diameters. We also developed simple models explaining the behavior we observed. Moreover, we introduced Kronecker graphs model with a rich set of properties, and developed algorithms for estimating its parameters. On the information cascade side we presented analyses of information propagation in large blog and recommendation networks, and developed scalable algorithms for early cascade detection.

The future plans for this thesis are to (a) analyze patterns of static and time evolving networks for anomaly detection and extrapolations, (b) build theories explaining the behavior and patterns we observe, and (c) build scalable tools for network analysis.

In the long run, outside the scope of this thesis, we would like to build tools for modeling the evolution of large networks both on a global scale and also on the micro-scale at the level of nodes or small communities. We want to study how information flows over the network and how local communities influence the global network and its evolution. Ideally, we want to bring these two views together, so that we can describe the evolution of the network as a whole, and at the same time also of its subparts.

References

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43.

Adar, E. and Adamic, L. A. (2005). Tracking information epidemics in blogspace. In *Web Intelligence*, pages 207–214.

Aiello, W., Chung, F., and Lu, L. (2000). A random graph model for massive graphs. In *STOC '00: Proceedings of the 32nd annual ACM symposium on Theory of computing*, pages 171–180.

Albert, R. and Barabasi, A.-L. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

Albert, R., Jeong, H., and Barabasi, A.-L. (1999). Diameter of the world-wide web. *Nature*, 401:130–131.

Anderson, R. M. and May, R. M. (2002). *Infectious diseases of humans: Dynamics and control.* Oxford Press.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications.* Hafner Press, 2nd edition.

Barabasi, A.-L., Ravasz, E., and Vicsek, T. (2001). Deterministic scale-free networks. *Physica A*, 299:559–564.

Bezáková, I., Kalai, A., and Santhanam, R. (2006). Graph model selection using maximum likelihood. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 105–112.

Bi, Z., Faloutsos, C., and Korn, F. (2001). The DGX distribution for mining massive, skewed data. In *KDD '01: Proceedings of the 6th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 17–26.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change in informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Blum, A., Chan, H., and Rwebangira, M. (2006). A random-surfer web-graph model. In *ANALCO '06: Proceedings of the 3rd Workshop on Analytic Algorithmics and Combinatorics*.

Bollobas, B. and Riordan, O. (2004). The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34.

Broder, A., Kumar, R., Maghoul1, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web: experiments and models. In *WWW '00: Proceedings of the 9th international conference on World Wide Web*.

Brown, J. J. and Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *The Journal of Consumer Research*, 14(3):350–362.

Carlson, J. M. and Doyle, J. (1999). Highly optimized tolerance: a mechanism for power laws in designed systems. *Physical Review E*, 60(2):1412–1427.

Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Computing Survey*, 38(1):2.

Chakrabarti, D., Leskovec, J., Faloutsos, C., Madden, S., Guestrin, C., and Faloutsos, M. (2007). Information survival threshold in sensor and p2p networks. In *INFOCOM '07: Proceedings of the 26th annual IEEE Conference on Computer Communications*.

Chakrabarti, D., Zhan, Y., and Faloutsos, C. (2004). R-mat: A recursive model for graph mining. In *SDM '04: SIAM Conference on Data Mining*.

Chevalier, J. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345.

Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882.

Cooper, C. and Frieze, A. (2003). A general model of web graphs. *Random Structures and Algorithms*, 22(3):311–335.

Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *KDD '01: Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. (2002). Pseudofractal scale-free web. *Physical Review E*, 65(6):066122.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Acadamy of Science*, 5:17–67.

Fabrikant, A., Koutsoupias, E., and Papadimitriou, C. H. (2002). Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *ICALP '02: Proceedings of the 29th International Colloquium on Automata, Languages, and Programming*, volume 2380.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. In *Proceedings of the National Academy of Sciences*, volume 99, pages 7821–7826.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., and Tomokiyo, T. (2005). Deriving marketing intelligence from online discussion. In *KDD '05: Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 419–428.

Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 3(12):211–223.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78:1360–1380.

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653.

Huberman, B. A. and Adamic, L. A. (1999). Growth dynamics of the world-wide web. *Nature*, 399:131.

Kempe, D., Kleinberg, J. M., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146.

Khuller, S., Moss, A., and Naor, J. (1999). The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45.

Kleinberg, J. M. (1999). The small-world phenomenon: an algorithmic perspective. Technical Report 99-1776, Cornell Computer Science Department.

Kleinberg, J. M., Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). The web as a graph: Measurements, models and methods. In *COCOON '99: Proceedings of the International Conference on Combinatorics and Computing*.

Krapivsky, P. L. and Redner, S. (2005). Network growth by copying. *Physical Review E*, 71(036118):036118.

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2003). On the bursty evolution of blogspace. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 568–576.

Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. (2000). Stochastic models for the web graph. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 57.

Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481–1493.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2006a). The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007a). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):2.

Leskovec, J., Chakrabarti, D., Kleinberg, J. M., and Faloutsos, C. (2005a). Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *PKDD '05: Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 133–145.

Leskovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636.

Leskovec, J. and Faloutsos, C. (2007). Scalable modeling of real graphs using kronecker multiplication. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*.

Leskovec, J., Horvitz, E., and Dumais, S. (2007b). Web projections: Learning from contextual subgraphs of the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*.

Leskovec, J., Kleinberg, J. M., and Faloutsos, C. (2005b). Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187.

Leskovec, J., Kleinberg, J. M., and Faloutsos, C. (2007c). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007d). Cost-effective outbreak detection in networks. In *Submitted to ACM KDD '07*.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. S., and Hurst, M. (2007e). Cascading behavior in large blog graphs. In *SDM '07: SIAM Conference on Data Mining*.

Leskovec, J., Singh, A., and Kleinberg, J. M. (2006b). Patterns of influence in a recommendation network. In *PAKDD '06: Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 380–389.

Li, L., Alderson, D., Doyle, J. C., and Willinger, W. (2005). Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523.

Milgram, S. (1967). The small-world problem. *Psychology Today*, 2:60–67.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251.

Montgomery, A. L. (2001). Applying quantitative marketing techniques to the internet. *Interfaces*, 30:90–108.

Nemhauser, G., Wolsey, L., and Fisher, M. (1978). An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351.

Newman, M. E. J., Forrest, S., and Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101.

Oliveira, J. G. and Barabasi, A. L. (2005). Human dynamics: The correspondence patterns of darwin and einstein. *Nature*, 437:1251.

Ostfeld, A., Uber, J. G., and Salomons, E. (2006). Battle of water sensor networks: A design challenge for engineers and algorithms. In *8th Symposium on Water Distribution Systems Analysis*.

Palmer, C. R., Gibbons, P. B., and Faloutsos, C. (2002). Anf: a fast and scalable tool for data mining in massive graphs. In *KDD '02: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., and Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211.

Ravasz, E. and Barabasi, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112.

Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B*, 4:131–134.

Resnick, P. and Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. In *The Economics of the Internet and E-Commerce*. Elsevier Science.

Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *KDD '02: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70.

Rogers, E. M. (1995). *Diffusion of Innovations*. Free Press, New York, fourth edition.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42:425–440.

Tauro, S. L., Palmer, C., Siganos, G., and Faloutsos, M. (2001). A simple conceptual model for the internet topology. In *GLOBECOM '01: Global Telecommunications Conference*, volume 3, pages 1667 – 1671.

Vazquez, A. (2001). Disordered networks generated by recursive searches. *Europhysics Letters*, 54(4):430–435.

Vazquez, A., Oliveira, J. G., Dezso, Z., Goh, K.-I., Kondor, I., and Barabasi, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127.

Wasserman, S., Faust, K., and Iacobucci, D. (1994). *Social Network Analysis : Methods and Applications*. Cambridge University Press.

Watts, D. J., Dodds, P. S., and Newman, M. E. J. (2002). Identity and search in social networks. *Science*, 296:1302–1305.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

Waxman, B. (1988). Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622.

Zipf, G. (1949). *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology.* Addison Wesley, Cambridge, Massachusetts.