# Information Propagation and Network Evolution on the Web

Mary McGlohon, Jure Leskovec, Christos Faloutsos*
Natalie Glance†, Matthew Hurst‡

## 1 Introduction

Using data gathered from blogs, this work seeks to understand the structure and formation of social networks, and the patterns of information propagation through these networks. Blogs have become an important medium of communication and information on the World Wide Web. Due to their accessible and timely nature, they are also an intuitive source for data involving the formation of social networks and the spread of information and ideas. By studying link patterns of existing entities and new arrivals to a blog network, we can infer the way in which social networks are formed. And, by examining linking patterns from one blog post to another, we can infer the way information spreads through a social network over the Web.

We seek to discover how information propagates through an existing network. Do trees representing the flow of information maintain certain structural properties? Does traffic in the blog network exhibit bursty and/or periodic behavior? After a topic becomes popular, how does interest die off – linearly, or exponentially? What models best exhibit such behavior?

We would also like to gain an understanding of how different entities in the social network function with regard to the propagation of information. Do some blogs act as hubs of information, often starting cascades of information to flow? Do certain subnetworks have different patterns of information propagation?

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
†Google, Pittsburgh, PA, USA
‡Microsoft Live Labs, Bellevue, WA, USA

### 1.1 Summary of findings and contributions

We note several temporal and topological observations in a blog network. Temporally, we note periodic behavior in traffic, and demonstrate that, surprisingly, post popularity drops off with a power law. Most topological network characteristics follow power laws: in-degree, out-degree, cascade size, and size of particular cascade shapes. We also produce a simple epidemiological model which captures most of the topological characteristics.

Delving further into the network entities, we use PCA to show that posts belonging to the same blog cluster into certain network behaviors, and that blogs in the same genre tend to cluster together based on their participation in different cascade shapes. In one specific case, we show that conservative blogs have more tree-like structures, while humorous blogs behave more like "stars".

## 2 Related Work

### 2.1 Blogs and social networks

Much work on modeling link behavior in large-scale on-line data has been done in the blog domain [1, 2, 22]. The authors note that, while information propagates between blogs, examples of genuine cascading behavior appeared relatively rare. This may, however, be due in part to the Web-crawling and text analysis techniques used to infer relationships among posts [2, 15]. Our work here differs in a way that we concentrate solely on the propagation of links, and do not infer additional links from text of the post, which gives us more accurate information.

There are several potential models to capture the structure of the blogosphere in particular, and of social networks in general. Work on information diffusion based on topics [15] showed that for some topics, their popularity remains constant in time ("chatter") while for other topics the popularity is more volatile ("spikes"). Authors in [22] analyze community-level behavior as inferred from blog-rolls – permanent links between "friend" blogs. Authors extended this work in [23] to analysis of several topological properties of link graphs in communities, finding that much behavior was characterized by "stars". Analysis based on thresholding as well as alternative probabilistic models of node activation is considered in the context of finding the most influential nodes in a network [18], and for viral marketing [26]. Such analytical work posits a known network, and uses the model to find the most influential nodes.

A number of generative models have been proposed for social networks in general [14, 19, 21, 35] Fitting static statistical models has also proved successful. They may be fit directly to the data, and parameters are estimated. A well-known class is exponential random graph, or $p*$, models. Based on Frank and Strauss' Markov graphs [12], $p*$ models are models defined by certain statistics of a graph, such as transitivity (if $a$ is a friend of $b$ and $c$ of $b$, $c$ is a friend of $a$), or triangles. The model places binary values on potential links, and parameters are then fit to empirical data. They may be used to define complicated dependency patterns [29].

### 2.2 Information cascades

Information cascades are phenomena in which an action or idea becomes widely adopted due to the influence of others, typically, neighbors in some network [5, 13, 14]. Cascades on random graphs using a threshold model have been theoretically analyzed [34]. Empirical analysis of the topological patterns of cascades in the context of a large product recommendation network is in [25] and [24].

As Carley addresses in [6], the diffusion of information and influence through a social network is greatly affected by the topology of the network. For this reason, it is important to address topological structure before studying cascades, which we will do in this work.

### 2.3 Virus propagation models in epidemiology

The study of epidemics offers powerful models for analyzing the spread of viruses. Our topic propagation model is based on the *SIS* (Susceptible-Infected-Susceptible) model of epidemics [3]. This is models flu-like viruses, where an entity begin as "susceptible", may become "infected" and infectious, and then heals to become susceptible again. A key parameter is the infection probability $\beta$, that is, the probability of a disease transmission in a single contact. Of high interest is the *epidemic threshold*, that is, the critical value of $\beta$, above which the virus will spread and create an epidemic, as opposed to becoming extinct. There is a huge literature on the study of epidemics on full cliques, homogeneous graphs, infinite graphs (see [16] for a survey), with recent studies on power-law networks [10] and arbitrary networks [33].

There have also been applications of these models in the domain of computer viruses. For example, Chen and Carley propose using state models

similar to SIR for modeling countermeasure propagation, in order to compete with virus propagation [7].

### 2.4 Self-similarity and heavy-tailed distributions

Self-similarity is often a result of heavy-tailed dynamics. Human interactions may be modeled with networks, and attributes of these networks often follow *power law* distributions [11]. Such distributions have a PDF (probability density function) of the form $p(x) \propto x^{\gamma}$, where $p(x)$ is the probability to encounter value $x$ and $\gamma$ is the exponent of the power law. In log-log scales, such a PDF gives a straight line with slope $\gamma$. 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 [36], with slope $1/(1 + \gamma)$. For $\gamma = -2$ we have the standard Zipf distribution, and for other values of $\gamma$ we have the generalized Zipf distribution.

Fitting heavy-tailed distributions is done in different ways. The method used in this work is done by taking histogram data, then taking the logarithm f both axes, and fitting a least-squares regression line to the log-log data. An $R^2$ coefficient greater than 0.95 suggests the data may be well-approximated with a power law distribution. Sometimes the tail is truncated, in this case because of artifacts of the data. One may also fit the empirical distribution, instead of the histogram data, with least-squares regression.

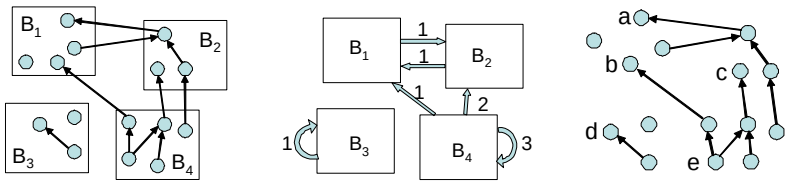
There are statistical methods of determining exactly which heavy-tailed distribution (if any) by which data are best approximated, such as maximum likelihood, and the Komolgorov-Smirnov test in Bayesian model selection. Detail of these methods may be found in work by Clauset, Shalizi, and Newman [8] and by Stouffer, Malmgren, and Amaral [30].

## 3 Preliminaries

In this section we introduce terminology and concepts regarding blog networks and information cascades.

### 3.1 Blogs and Cascades

Blogs (weblogs) are web sites that are updated on a regular basis. Blogs have the advantage of being easy to access and update, and have come to serve a variety of purposes. Often times individuals use them for online diaries and social networking; other times news sites have blogs for timely




(a) Blogosphere (b) Blog network (c) Post network

Figure 1: The model of the blogosphere (a). Squares represent blogs and circles blog-posts. Each post belongs to a blog, and can contain hyperlinks to other posts and resources on the web. We create two networks: a weighted blog network (b) and a post network (c). Nodes $a, b, c, d$ are *cascade initiators*, and node $e$ is a *connector*.

stories. Blogs are composed of posts that typically have room for comments by readers – this gives rise to discussion and opinion forums that are not possible in the mass media. Also, blogs and posts typically link each other, as well as other resources on the Web. Thus, blogs have become an important means of transmitting information. The influence of blogs was particularly relevant in the 2004 U.S. election, as they became sources for campaign fund-raising as well as an important supplement to the mainstream media [1]. Understanding the ways in which information is transmitted among blogs is important to developing concepts of present-day communication.

We model two graph structures emergent from links in the blogosphere, which we call the *Blog network* and the *Post network*. Figure 1 illustrates these structures. Blogosphere is composed of blogs, which are further composed of posts. Posts then contain links to other posts and resources on the web.

From Blogosphere (a), we obtain the Blog network (b) by collapsing all links between blog posts into directed edges between blogs. A blog-to-blog edge is weighted with the total number of links where a post in source blog points to a post in destination blog. From the Blog network we can infer a social network structure, under the assumption that blogs that are "friends" link each other often.

In contrast, to obtain the Post network (c), we ignore the posts' parent blogs and focus on the link structure. Associated with each post is the date of the post, so we label the edges in Post network with the date difference $\Delta > 0$ between the source and the destination posts. Let $t_u$ and $t_v$ denote post times of posts $u$ and $v$, where $u$ links to $v$, then the link time $\Delta = t_u - t_v$.

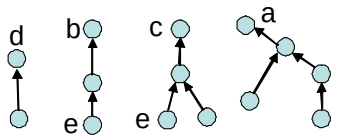



Figure 2: Cascades extracted from Figure 1. Cascades represent the flow of information through nodes in the network. To extract a cascade we begin with an initiator with no out-links to other posts, then add nodes with edges linking to the initiator, and subsequently nodes that link to any other nodes in the cascade.

From the Post network, we extract information cascades, which are induced subgraphs by edges representing the flow of information. A cascade (also known as conversation tree) has a single starting post called the *cascade initiator* with no out-links to other posts (e.g. nodes $a, b, c, d$ in Figure 1(c)). Posts then join the cascade by linking to the initiator, and subsequently new posts join by linking to members within the cascade. Figure 2 gives a list of cascades extracted from Post network in Figure 1(c). Since a link points from the follow-up post to the existing (older) post, influence propagates following the reverse direction of the edges.

We define a *non-trivial* cascade to be a cascade containing at least two posts. Therefore, a *trivial cascade* is an isolated post. Figure 2 shows all non-trivial cascades in Figure 1(c), but not the two trivial cascades. Cascades form two main shapes, which we refer to as *stars* and *chains*. A star occurs when a single center post is linked by several other posts, but the links do not propagate further. This produces a wide, shallow tree. Conversely, a chain occurs when a root is linked by a single post, which in turn is linked by another post. This creates a deep tree that has little breadth. As we will later see most cascades are somewhere between these two extreme points.

### 3.2 Principal component analysis

Given many vectors in $D$-dimensional space, how can one visualize them, when the dimensionality $D$ is high? This is exactly where principal component analysis (PCA) helps. PCA will find the optimal 2-dimensional plane to project the data points, maintaining the pair-wise distances as best as possible. PCA is even more powerful than that: it can give us a sorted list of directions ("principal components") on which we can project. See [17] or [20] for more details.

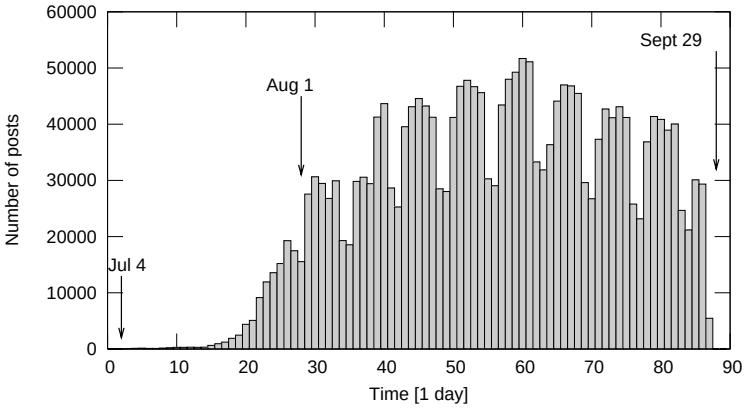



Figure 3: Number of posts by day over the three-month period.

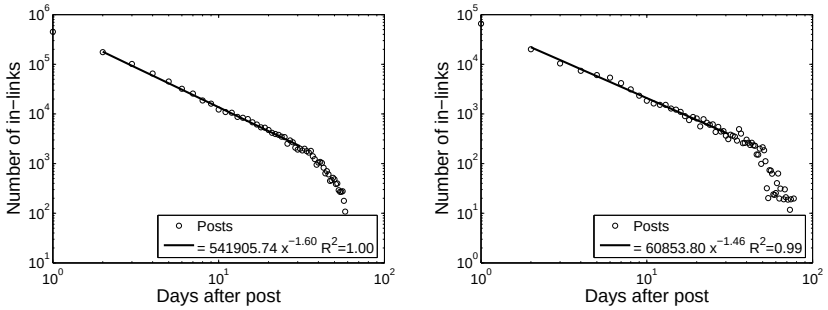
Traditional PCA assumes a Gaussian distribution. However, the data we will be dealing with is heavy-tailed, and may be approximated with a log-normal distribution. Therefore, in our experiments we will choose to take the logarithm of each value– if the data is close to a log-normal this will transform it to a normal distribution and thereby behaves well with PCA. A similar method of improving PCA's performance by normalizing data is TF-IDF (term frequency– inverse document frequency), detailed in [28].

A generalized, more robust, version is exponential family PCA. PCA may be interpreted as fitting maximum likelihood of parameters $\Theta$ in a multivariate unit Gaussian. Using generalized linear models, this approach extends that interpretation of PCA to estimate parameters for any exponential family model, and not simply Gaussian. This approach is particularly valuable for integer-valued or binary-valued data, which may be better approximated with Poisson or Bernoulli distributions. Details may be found in [9]. We choose traditional PCA for this work, and believe this approach is suitable since the data may be closely approximated with a log-normal.

# 4 Observations and Experiments

## 4.1 Temporal patterns

Traffic in the blogosphere is not uniform. As Figure 3 illustrates, there is a seven-day periodicity. Posting and blog-to-blog linking patterns tend to have a *weekend effect*, with frequency sharply dropping off at weekends. In




All posts Only Monday posts

Figure 4: Number of in-links vs. the days after the post in log-linear scale, after removing the day-of-the week effects. Power law fits to the data, based on reasonable lookahead of 30 days, produce exponents $-1.6$ and $-1.46$.

Figure 3 we plot the number of posts per day over the span of our dataset.

Next, we examine how a post's popularity grows and declines over time. 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. By aggregating over a large set of posts we obtain a more general pattern.

However, the weekend effect creates abnormalities in the plots we must account for. We smooth the in-link plots by applying a weighting parameter to the plots separated by day of week. For each delay $\Delta$ on the horizontal axis, we estimate the corresponding day of week $d$, and we prorate the count for $\Delta$ by dividing it by $l(d)$, where $l(d)$ is the percent of blog links occurring on day of week $d$. This simulates a popularity drop-off that might occur if posting and linking behavior were uniform throughout the week.

We fit the power-law distribution with a cut-off in the tail (bottom row). We fit on 30 days of data, as most posts in the graph have complete in-link patterns for the 30 days following publication. We performed the fitting over all posts and for all days of the week separately, and found a stable power-law exponent of around $-1.5$, 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 [4] – 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.

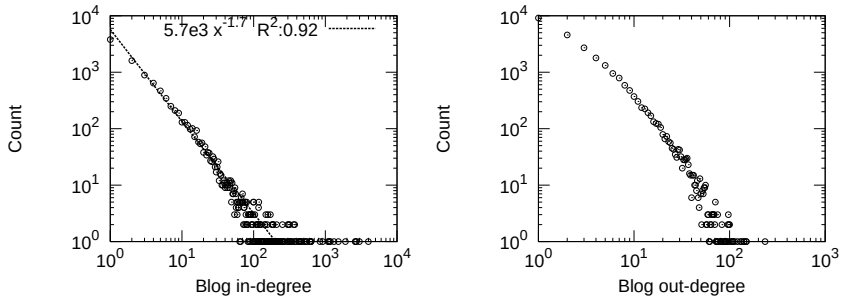



Figure 5: Blog network in- and out-degree distributions.

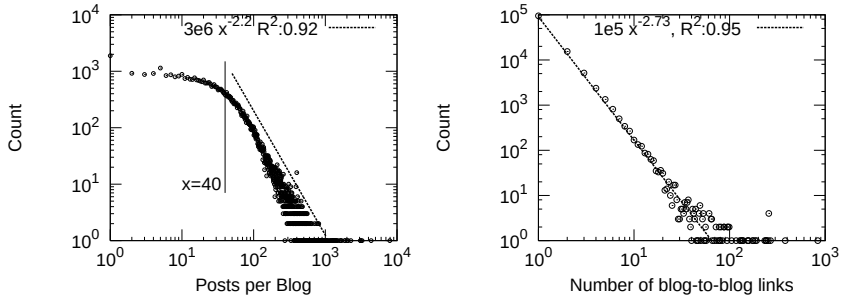



Figure 6: Distribution of the number of posts per blog (a); Distribution of the number of blog-to-blog links, i.e. the distribution over the Blog network edge weights (b).

## 4.2 Blog network and Post network topology

The first graph we consider topologically is the Blog network. As illustrated in Figure 1(c), every node represents a blog and there is a weighted directed edge between blogs $u$ and $v$, where the weight of the edge corresponds to the number of posts from blog $u$ linking to posts at blog $v$. Connectivity-wise, half of the blogs belong to the largest connected component and the other half are isolated blogs.

We show the in- and out-degree distribution in Figure 5. Notice both follow a heavy-tailed distribution. The number of posts per blog, as shown in Figure 6(a), also follows a heavy-tailed distribution. The deficit of blogs with low number of posts and the knee at around 40 posts per blog can be explained by the fact that we are using a dataset biased towards active

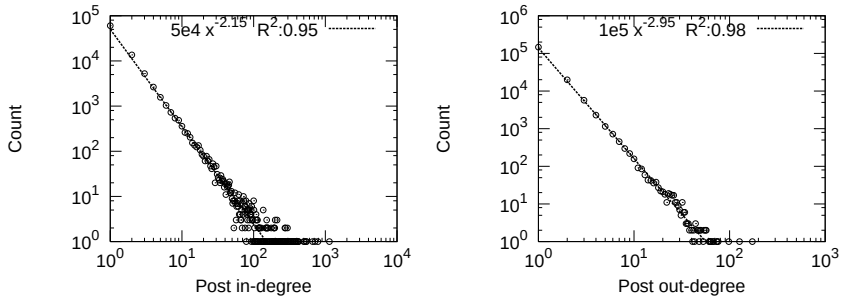



Figure 7: Post network in- and out-degree distribution.

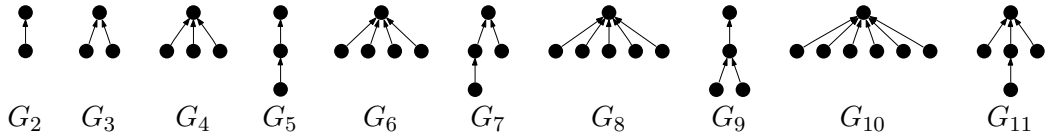



Figure 8: Common cascade shapes ordered by frequency. Cascade with label $G_r$ has the frequency rank $r$.

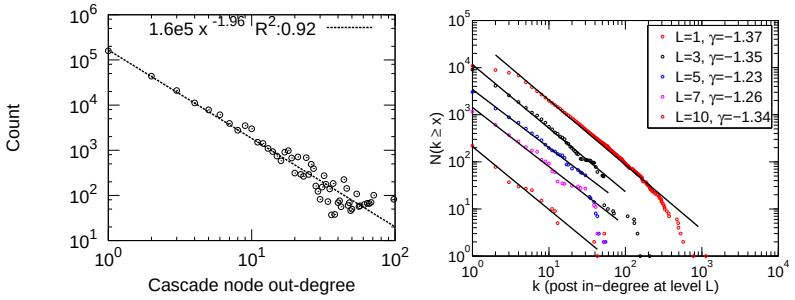
blogs. However, our sample still maintains the power law in the number of blog-to-blog links as shown in 6(b).

In contrast to Blog network the Post network is very sparsely connected. 98% of the posts are isolated, and the largest connected component accounts for $106,000$ nodes, while the second largest has only 153 nodes. Figure 7 shows the in- and out-degree distributions of the Post network which, not surprisingly, follow a power law.

### 4.3 Information propagation through cascades

We are especially interested in how information propagates, and this phenomenon is illustrated by cascades. Given the Post network we extracted all information cascades using the following procedure. We found all cascade initiator nodes, i.e. nodes that have zero out-degree, and started following their in-links. This process gives us a directed acyclic graph with a single root node. To obtain the examples of the common shapes and count their frequency we used algorithms described in [25]. We find a total of $2,092,418$ cascades.

We give examples of common Post network cascade shapes in Figure 8. A node represents a post and the influence flows from the top to the bot-




(a) Out-degree (b) In-degree at level $L$

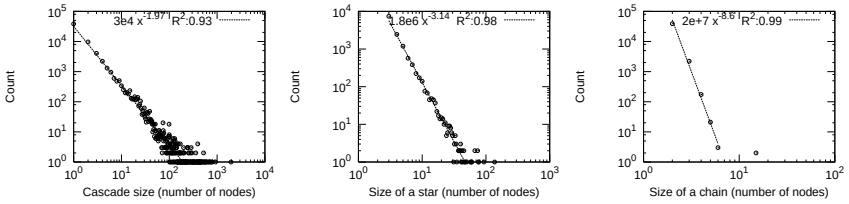
Figure 9: Out-degree distribution over all cascades extracted from the Post network (a), and the in-degree distribution at level $L$ of the cascade (b). Note all distributions are heavy tailed and the in-degree distribution is remarkably stable over the levels.

tom. Cascades tend to be wide and not too deep– stars and shallow bursty cascades are the most common type of cascades.

We next examine the general cascade behavior by measuring and characterizing the properties of real cascades. First, we observe the degree distributions of the cascades. This means that from the Post network we extract all the cascades and measure the overall degree distribution. Essentially we work with a *bag of cascades*, where we treat a cascade as separate disconnected sub-graph in a large network. Similar to other networks, in- and out-degree distribution of the bag of cascades follow power laws with exponents of -2.2 and -1.92, respectively (Figure 9). Further examination showed that the in-degree exponent is stable and does not change much given level $L$ in the cascade (a node is at level $L$ if it is $L$ hops away from the cascade initiator). 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 next ask: what distribution do cascade sizes follow? Does the probability of observing a cascade on $n$ nodes decreases exponentially with $n$? We examine the *Cascade Size Distributions* over the bag of cascades extracted from the Post network. We consider three different distributions: a distribution over all cascade sizes, and separate size distributions of star and chain cascades. We chose stars and chains since they are well defined and given the number of nodes in the cascade, there is no ambiguity in the topology of a star or a chain.

Figure 10 gives the Cascade Size Distribution plots. Notice all follow a




(a) All cascades (b) Star cascade (c) Chain cascade

Figure 10: Size distribution over all cascades (a), only stars (b), and chains (c). They all follow heavy tailed distributions with increasingly steeper slopes.

heavy-tailed distribution, with slopes $\approx -2$ overall (Figure 10(a)). So the probability of observing a cascade on $n$ nodes follows a Zipf distribution: $p(n) \propto n^{-2}$. Stars have the power-law exponent $\approx -3.1$ (Figure 10(b)), and chains are small and rare and decay with exponent $\approx -8.5$ (Fig. 10(c)).

### 4.4 Cascade generation model

We present 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) are the conversations in general on the blogosphere.

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 being a graph created by the spread of the 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 $\beta$ that measures how infectious are the posts on the blogosphere. Our model is very similar to the SIS (susceptible – infected – susceptible) model from the epidemiology [16].

Next, we describe the model. Each blog is in one of two states: *infected* or *susceptible*. If a blog is in the infected state this means that the blogger just posted a post, and the blog now has a chance to spread its influence. Only blogs in the susceptible (not infected) state can get infected. When a blog successfully infects another blog, a new node is added to the cascade,

and an edge is created between the node and the source of infection. The source immediately recovers, i.e. a node remains in the infected state only for one time step. This gives the model ability to infect a blog multiple times, which corresponds to multiple posts from the blog participating in the same cascade.

More precisely, a single cascade of the *Cascade generation model* is generated by the following process.

(i) Uniformly at random pick blog $u$ in the Blog network as a starting point of the cascade, set its state to *infected*, and add a new node $u$ to the cascade graph.

(ii) Blog $u$ that is now in infected state, infects each of its uninfected directed neighbors in the Blog network independently with probability $\beta$. Let $\{v_1, \ldots, v_n\}$ denote the set of infected neighbors.

(iii) Add new nodes $\{v_1, \ldots, v_n\}$ to the cascade and link them to node $u$ in the cascade.

(iv) Set state of node $u$ to not infected. Continue recursively with step (ii) until no nodes are infected.

We make a few observations about the proposed model. First, note that the blog immediately recovers and thus can get infected multiple times. Every time a blog gets infected a new node is added to the cascade. This accounts for multiple posts from the blog participating in the same cascade. Second, we note that in this version of the model we do not try to account for topics or model the influence of particular blogs. We assume that all blogs and all conversations have the same value of the parameter $\beta$. Third, the process as describe above generates cascades that are trees. This is not big limitation since we observed that most of the cascades are trees or tree-like. In the spirit of our notion of cascade we assume that cascades have a single starting point, and do not model for the collisions of the cascades.

### 4.4.1 Validation of the model

We validate our model by extensive numerical simulations. We compare the obtained cascades towards the real cascades extracted from the Post network. We find that the model matches the cascade size and degree distributions.

We use the real Blog network over which we propagate the cascades. Using the Cascade generation model we also generate the same number of

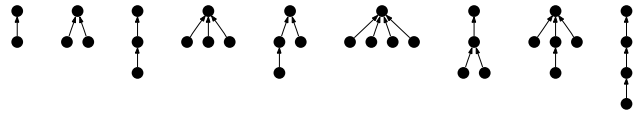


Figure 11: Top 10 most frequent cascades as generated by the Cascade generation model. Notice similar shapes and frequency ranks as in Figure 8.

cascades as we found in Post network ($\approx$ 2 million). We tried several values of $\beta$ parameter, and at the end decided to use $\beta = 0.025$. This means that the probability of cascade spreading from the infected to an uninfected blog is 2.5%. We simulated our model 10 times, each time with a different random seed, and report the average. We believe that since there was little variance in the behavior of the Cascade generation model 10 runs was sufficient.

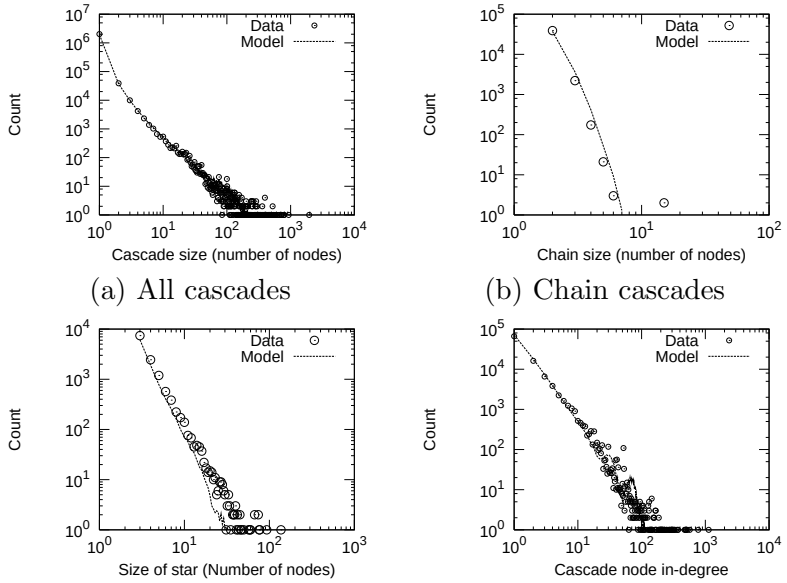
First, we show the top 10 most frequent cascades (ordered by frequency rank) as generated by the Cascade generation model in Figure 11. Comparing them to most frequent cascades from Figure 8 we notice that top 7 cascades are matched exactly (with an exception of ranks of $G_4$ and $G_5$ swapped), and rest of cascades can also be found in real data.

Next, we show the results on matching the cascade size and degree distributions in Figure 12. We plot the true distributions of the cascades extracted from the Post network with dots, and the results of our model are plotted with a dashed line. We compare four properties of cascades: (a) overall cascade size distribution, (b) size distribution of chain cascades, (c) size distribution of stars, and (d) in-degree distribution over all cascades.

Notice a very good agreement between the reality and simulated cascades in all plots. The distribution over of cascade sizes is matched best. Chains and stars are slightly under-represented, especially in the tail of the distribution where the variance is high. The in-degree distribution is also matched nicely, with an exception of a spike that can be attributed to a set of outlier blogs all with in-degree 52. Note that cascades generated by the Cascade generation model are all trees, and thus the out-degree for every node is 1.

### 4.4.2 Exponential random graph ($p^*$) model

It is worth noting that exponential random graphs would be another option for fitting and understanding the way in which cascades come about. As described in Section 2, $p^*$ models fit parameters for likelihoods of certain graph structures, such as reciprocal links and triangles. In applying $p^*$ to




(a) All cascades (b) Chain cascades

(c) Star cascades (d) In-degree distribution

Figure 12: 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.

our data, we would like to fit the nontrivial cascade shapes identified– such as stars and chains of different sizes, as well as more complex structures.

To perform fitting, we would follow the steps detailed in Robins et. al ([27]). We assume each tie is a random variable, develop hypotheses about the interdependencies among RV ties, estimate parameters in the model, and evaluate the model.

In our case, we have 2 million nodes, each representing one post. First, we consider each possible network tie as a binary random variable– for any posts $p_i$ and $p_j$, we will have $X_{ij}$ to respresent the existence of a link from $p_i$ to $p_j$. The observed value of $X_{ij}$ is represented as $x_{ij}$, and we can represent the entire network as an adjacency matrix $\mathbf{X}$. In our case $\mathbf{X}$ will not be symmetric, in fact it will be strictly asymmetric, considering there are no cycles due to timing of posts.

The next step is to make a hypothesis about the interdependencies among network variables. For $p^*$, an edge variable $X_{ij}$ may be dependent on

other edge variables, or dependent on node attributes. In random graphs, ties are independent of each other– the existence of a link from $p_i$ to $p_j$ is independent of a link from $p_i$ to $p_k$, etc. However, we might choose the hypothesis that a post $p_k$ is more likely to link to a post $p_i$ if there is some $p_j$ such that $X_{ji} = 1$. We might also choose to add that if $X_{ji} = 1$ then $E(X_{ki}) < E(X_{kj})$, a hypothesis suggesting that "stars" are more inclined to form than "chains".

Once we decide on hypotheses to examine, we build up the model to test these hypotheses, and the dependence assumptions imply the general model. In the hypotheses we have identified a number of *configurations* of interest, such as two-stars, three-stars, two-chains, three-chains, etc. We will denote the set of configurations as $\mathbf{C}$. We consider each parameter in the model to be the probability of a configuration in the network, for instance, a two-star or three-star, where the count of each configuration in $c \in \mathbf{C}$ is represented by statistic $z_c$, and the parameter is a corresponding function $\theta_c z_c$. The full specification of the model will then represent a distribution of random graphs which are "built up" from the localized patterns. There may be a parameter related to the likelihood of a two-star occurring (denoted $\theta_1 z_1$), a three-star occurring ($\theta_2 z_2$), etc. We would then be able to represent the probability of the entire graph $\mathbf{X}$ taking on values $\mathbf{x}$ as:

$$P(\mathbf{X} = \mathbf{x}) = \frac{exp(\sum_{c \in \mathbf{C}} \theta_c z_c(\mathbf{x}))}{\kappa(\Theta)}$$

where $z_c(\mathbf{x})$ is the number of times configuration $c$ occurs in candidate graph $\mathbf{x}$ and $\kappa(\Theta)$ is a normalizing constant that ensures that the probabilities of all possible graphs sum to one. The model thus represents a distribution of random graphs made up of the configurations of interest. (Note this is after taking *homogeneity constraints* into account, where instead of using a different parameter $\lambda_A$ for each instance $A$ of a configuration $c$, we equate each $\lambda_A$ with $\theta_c$, and use statistic $z_c(\mathbf{x})$ as a count [27]). For instance, $z_c$ for a star including four nodes (a "3-star", for a root and three followers) would be $\sum_{i,j,k,l} x_{ij} x_{ik} x_{il}$, and the statistic for counts of a chain including four nodes would be $\sum_{i,j,k,l} x_{ij} x_{jk} x_{kl}$.

We would then estimate parameter values through maximum likelihood. However, it becomes clear that such form of the model is intractable because of the normalizing constant $\kappa(\Theta)$. As described in Strauss and Ikeda [31], we can instead use a logit version of $p^*$. The logit model takes the form:

$$\omega_{ij} = log(\frac{P(X_{ij} = 1 | \mathbf{X}^c_{ij})}{P(X_{ij} = 0 | \mathbf{X}^c_{ij})} = \Theta[\mathbf{z}(\mathbf{x}^+_{ij}) - \mathbf{z}(\mathbf{x}^-_{ij})])$$

That is, the log odds ratio for an edge $X_{ij}$ conditioned on the values of the adjacency matrix besides $X_{ij}$, is the dot product of the parameters $\Theta$ and the difference of the statistics $\mathbf{z}$ when the value of $X_{ij}$ changes from 1 to 0.

From this, parameters are estimated by finding the *maximum pseudo-likelihood estimator* with respect to pseudolikelihood function:

$$PL(\Theta) = \prod_{i \neq j} P(X_{ij} = 1 | \mathbf{X}_{ij}^c)^{x_{ij}} P(X_{ij} = 0 | \mathbf{X}_{ij}^c)^{1 - x_{ij}}$$

It was proved that this can be estimated using logistic regression [31]. One finds the maximum likelihood fit of logistic reression to the model $\omega_{ij}$ above, easily implemented as an iteratively reweighted Gauss-Newton least squares procedure. This is easy to estimate, but the properties of the estimator are not as well understood; so Monte Carlo techniques may also be used [27].

An advantage of $p^*$ with respect to our data is that it can generate the cascade shape counts exactly. It is also flexible– for instance, if one desires to count $G_{11}$ in Figure 8 as both a chain and a star when estimating parameters, the model may be fitted accordingly.

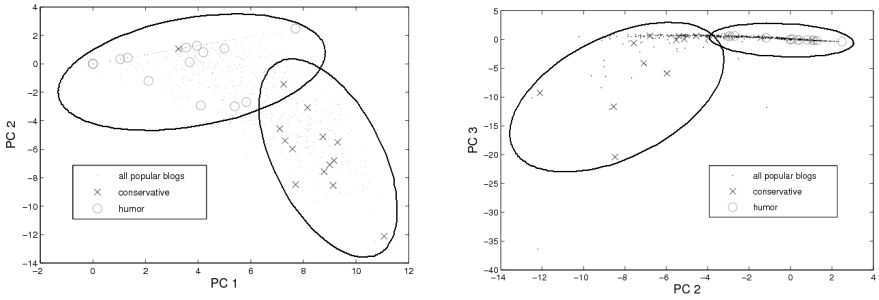
A potential shortcoming of the $p^*$ model is that it is not generative in the same way the Cascade generation model is: it does not directly provide an intuition for *how* cascades may formed on a low level. However, the estimated values are for the purposes of observing overall tendencies toward certain cascade properties (such as star tendency, chain tendency, or cascade size), and may provide valuable information toward building or checking a generative model. For instance, instead of finding $\beta$ by what seems like the "best count", we could instead find the parameter more precisely: with each simulated graph for a given $\beta$, we could plug the graph into the $p^*$ model and get a likelihood statistic that indicates how well the simulation on $\beta$ fits into what $p^*$ would predict.

## 4.5 Network Entities

Finally, we would like to gain an understanding of how different entities in the social network function with regard to the propagation of information.

### 4.5.1 Clustering blogs by Cascade Shapes

Our first experiments involved performing PCA on a large, sparse matrix where rows represented blogs and columns represented different types of




(a) First vs. second PC (b) Second vs. third PC

Figure 13: Principal components for blogs by CascadeType labeled by topic. PC's were generated by analyzing a matrix of blogs by counts of cascade types. Note that there is a clear separation between conservative blogs *(*represented by red crosses*)*, and humorous blogs *(*represented with by circles*)*, both on axes of the first and second PC *(*a*)*, and on axes of the second and third PC *(*b*)*. Ovals indicate the main clusters

cascades. Each entry was a count, and in order to reduce the variance, we took the log of each count. Our dataset consisted of $44,791$ blogs with $8,965$ cascade types.

It was of interest to impose social networks upon the blogs, based on what topics the blogs tended to focus on. We hand-classified a sample of the blogs in the data by topic, and found that we could often separate communities based on this analysis. For the purposes of visualization we chose to focus on two of the larger communities, politically conservative blogs and "humorous" blogs (such as blogs for different web-comics and humorists). Figure 13(a) shows these blogs plotted on the first two principal components, and Figure 13(b) shows them plotted on the second and third principal components. Ovals are drawn around the main clusters. We notice a distinct separation between the conservative community and the humor community; this means that the two communities engage in very different conversation patterns.

Upon closer analysis, we find this is the case because conservative blogs tend to form deep, chainlike graphs whereas the humorous blogs form stars. Some similar observations may be made for other communities; we used these two because they were the most distinct. This result shows that blog communities tend to follow different linking patterns. We believe that by looking at a blog's cascade types that one can better make inferences about

what community a blog might belong to.

Furthermore, the number of trivial cascades that a blog participates in–that is, its number of solitary posts with no in- or out-links, may be a key indicator of its community. Removing the trivial cascades caused the clusters to become less clear, which indicates that these trivial cascades still play a significant role in the inferences one can make about that blog.

### 4.5.2 Clustering posts based on features

We next sought to find how posts themselves behave. In order to do this, we performed PCA on a 6-column matrix. Each row represented a post, while the columns were as follows:

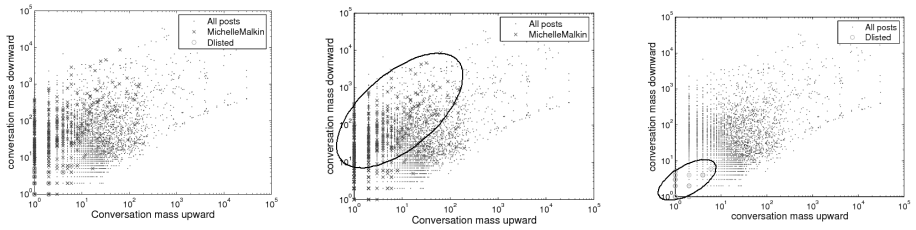
- Number of in-links
- Number of out-links
- Conversation mass upwards
- Conversation mass downwards
- Depth upwards
- Depth downwards

There were $6,666,188$ posts in the dataset. When we ran PCA, we found that the major two components that determined the blog's place in this space were conversation mass upwards and downwards. Therefore, we also plotted the posts on the two axes of conversation mass upwards and conversation mass downwards (See Figure 14. To illustrate, we have plotted all posts, with special markers for two distinct popular blogs, Dlisted[1] and MichelleMalkin.[2] We have circled the main clusters in the plots. Notice that while Dlisted and MichelleMalkin points overlap, their clusters are centered differently. The mean and variance of these clusters can serve as another viewpoint into the profile of a blog.

Thus, we observe that Posts within a blog tend to take on common network characteristics, which may serve as another means of classification. Individual posting patterns may serve as another way of clustering blogs, because different blogs maintain different posting patterns.

[1] `dlisted.blogspot.com`, a celebrity gossip blog.

[2] `www.MichelleMalkin.com`, a politically conservative blog.




(a) All posts (b) MichelleMalkin (c) Dlisted

Figure 14: Conversation mass for posts, an aspect of PostFeatures6. The top figure shows the Dlisted and MichelleMalkin clusters superimposed over points for all posts. The next two show the clusters separately, superimposed on all blog points for reference. Ovals are drawn around the main clusters. Note that while there is overlap between posts features of two blogs, they have different centers. This tells us that different blogs maintain different means and variances in conversation masses

# 5 Discussion

Our finding that the the popularity of posts drops off with a power law distribution is interesting since intuition might lead one to believe that people would "forget" a post topic in an exponential pattern. However, since linking patterns are based on the behaviors of individuals over several instances, much like other real-world patterns that follow power laws such as traffic to Web pages and scientists' response times to letters [32], it is reasonable to believe that a high number of individuals link posts quickly, and later linkers fall off with a heavy-tailed pattern.

Our findings have potential applications in many areas. One could argue that the conversation mass metric, defined as the total number of posts in all conversation trees below the point in which the blogger contributed, summed over all conversation trees in which the blogger appears, is a better proxy for measuring influence. This metric captures the mass of the total conversation generated by a blogger, while number of in-links captures only direct responses to the blogger's posts.

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 Fig. 8). On the other hand 53% of posts from a political blog MichelleMalkin were cascade initiators.

But the cascade here were deeper and generally larger (e.g., $G_{117}$ in Fig. 8) than those of BoingBoing.

The methods chosen for clustering were decided mainly for simplicity, as the main goal was to present ideas for some blog characterization. For CascadeType and PostFeatures6 we ran PCA after taking the log counts. There are other methods available for reducing variance, however, we chose log for the sake of simplicity. It may be of interest to use different forms of TF-IDF, a method often used in text mining. A description of TF-IDF is provided in [28].

We have analyzed many characteristics of blogs, based on conversation patterns, post features, and post patterns over time. From this basis, given a blog, we can infer a number of things about that blog based on these metrics.

# 6 Conclusion

We analyzed one of the largest available collections of blog information, trying to find how blogs behave and how information propagates through the blogosphere. We studied two structures, the "Blog network" and the "Post network". Our findings are summarized as follows:

*Temporal Patterns:* The decline of a post's popularity follows a power law, rather than an exponential dropoff as might be expected. The slope is ≈-1.5, the slope predicted by a very recent theory of heavy tails in human behavior [32].

*Topological Patterns:* Almost any metric we examined follows a power law: size of cascades, size of blogs, in- and out-degrees. Finally, stars and chains are basic components of cascades, with stars being more common. Most cascades are tree-like.

*Generative model:* Our idea is to reverse-engineer the underlying social network of blog-owners, and to treat the influence propagation between blog-posts as a flu-like virus, that is, the SIS model in epidemiology. Despite its simplicity, our model generates cascades that match very well the real cascades with respect to in-degree distribution, cascade size distribution, and popular cascade shapes. The model achieved this accuracy with a constant infectiousness value of $\beta$ and by weighting the closeness of linked blogs equivalently, even if they are linked multiple times.

*Characterizing blogs:* We have also made several observations on what sort of features best characterize blogs in a network. We made some observations about cascade types. First, we note that the cascade types that blogs participate may suggest to which community it belongs ('humor', 'con-

servative', etc.) . Second, the number of trivial (singleton) cascades that a blog uses is a major indicator of cascade type. We can characterize blogs based on their general network characteristics, and observed that blogs tend to have posts that cluster together with respect to post features.

Future work abounds, because the blogosphere is an extremely rich dataset, with multiple facets. Future research could try to include the content of the posts, to help us find even more accurate patterns of influence propagation. Another direction is to spot anomalies and link-spam attempts, by noticing deviations from our patterns.

## References


[1] L. A. Adamic and N. Glance. 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, 2005.

[2] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 207–214, Washington, DC, USA, 2005. IEEE Computer Society.

[3] N. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 1975.

[4] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207, 2005.

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

[6] K. M. Carley. On the evolution of social and organizational networks. *Research in the Sociology of Organizations*, 16(Special issue on Networks In and Around Organ):3–30, 1999.

[7] L.-C. Chen and K. M. Carley. The impact of countermeasure propagation on the prevalence of computer viruses. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(2):823–833, 2004.

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, 2007.

[9] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[10] V. M. Equiluz and K. Klemm. Epidemic threshold in structured scale-free networks. *arXiv:cond-mat/02055439*, May 21 2002.

[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *SIGCOMM*, pages 251–262, Aug-Sept. 1999.

[12] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

[13] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

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

[15] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, 2004.

[16] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653, 2000. http://www.math.rutgers.edu/ leenheer/hethcote.pdf.

[17] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[18] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03*, 2003.

[19] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The Web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science*, 1627:1–17, 1999.

[20] F. Korn, H. Jagadish, and C. Faloutsos. Efficiently supporting ad hoc queries in large datasets of time sequences. *ACM SIGMOD*, pages 289–300, May 13-15 1997.

[21] P. Krapivsky and S. Redner. Network growth by copying. *Physical Review E*, 71:036118, 2005.

[22] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[23] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowedge Discover and Data Mining*, pages 611–617, New York, 2006.

[24] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 228–237, New York, NY, USA, 2006. ACM Press.

[25] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

[26] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing, 2002.

[27] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. A workshop on exponential random graph models for social networks, 2005.

[28] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, Cornell University, Ithaca, NY, USA, 1987.

[29] T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, December 2006.

[30] D. B. Stouffer, R. D. Malmgren, and L. A. N. Amaral. Log-normal statistics in e-mail communication patterns, 2006.

[31] D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of American Statistical Association*, 85(409):204–212, March 1990.

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

[33] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *SRDS*, pages 25–34, 2003.

[34] D. J. Watts. A simple model of global cascades on random networks. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 99, pages 5766–5771, April 30 2002. http://www.jstor.org/view/00278424/sp020038/02x3936j/0.

[35] D. J. Watts and S. H. Strogatz. *Collective dynamics of 'small world' networks*, chapter 4.2, pages 301–303. Princeton University Press, 2006.

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

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. IIS-0209107, SENSOR-0329549, EF-0331657, IIS-0326322, IIS-0534205, and also by the Pennsylvania Infrastructure Technology Alliance (PITA). Additional funding was provided by a generous gift from Hewlett-Packard. Mary McGlohon was partially supported by a National Science Foundation Graduate Research Fellowship.