Information Diffusion Through Blogspace

D. Gruhl
IBM Research, Almaden
dgruhl@us.ibm.com

R. Guha
IBM Research, Almaden
rguha@us.ibm.com

David Liben-Nowell
Laboratory for Computer Science, MIT
dln@theory.lcs.mit.edu

A. Tomkins
IBM Research, Almaden
tomkins@almaden.ibm.com

ABSTRACT

We study the dynamics of information propagation in environments of low-overhead personal publishing, using a large collection of WebLogs over time as our example domain. We characterize and model this collection at two levels. First, we present a macroscopic characterization of topic propagation through our corpus, formalizing the notion of long-running "chatter" topics consisting recursively of "spike" topics generated by outside world events, or more rarely, by resonances within the community. Second, we present a microscopic characterization of propagation from individual to individual, drawing on the theory of infectious diseases to model the flow. We propose, validate, and employ an algorithm to induce the underlying propagation network from a sequence of posts, and report on the results.

1. INTRODUCTION

Over the course of history, the structure of societies and the relations between different societies have been shaped to a great extent by the flow of information in them [11]. More recently, over the last 15–20 years, there has been interest not just in observing these flows, but also in influencing and creating them. Doing this requires a deep understanding of the macro- and micro-level structures involved, and this in turn has focused attention on modeling and predicting these flows.

The mainstream adoption of the Internet and Web has changed the physics of information diffusion. Until a few years ago, the major barrier for someone who wanted a piece of information to spread through a community was the cost of the technical infrastructure required to reach a large number of people. Today, with widespread access to the Internet, this bottleneck has largely been removed. In this context, *personal publishing* modalities such as weblogs have become prevalent.

Weblogs, unlike earlier mechanisms for spreading information at the grassroots level, offer the opportunity for direct, frequent, and low-cost observation of information flow at the individual level. This in turn enables applications that were not previously possible. Given the huge sums of effort and money spent by corporations and political organizations to spread their message, timely feedback and monitoring are vital to maximizing the impact of a marketing, political, or other campaign. On the other side, users are inundated with organizations clamoring for their attention. We want to leverage the blogging community to identify newsworthy events, as evidenced by spikes in postings of the relevant communities.

We are interested in the dynamics of information propagation in environments of low-overhead personal publishing, such as web pages, Weblogs, bulletin boards, and netnews. We focus in this paper on *Blogspace*, the space of all weblogs. Of course, personal publishing doesn't occur in isolation. It is influenced by, and influences, the older more mainstream media sources. Thus, in our analysis, we include both Weblog postings and news articles from sources such as Reuters and the AP Newswire.

An obvious course of analysis of blogspace would be based on the link structure manifested in blogrolls and such. We posit that blogspace exhibits distinct structures when examined at different temporal granularities. At a coarse granularity, we find the kind of structure described in [22]. Our focus is not so much on this structure of blogspace as in the diffusion of information, as reflected in who influences whom, which is a much more dynamic structure. In doing this, we find that traditional media sources such as Reuters and AP (who do not typically appear in blogrolls) still have an enormous influence. Thus, we believe that our study applies more generally to the diffusion of information in environments of personal publishing and not just to blogspace.

There are many dimensions along which information diffusion can be characterized. In this paper, we explore the following:

Topics: We are interested in first identifying the set of postings that are *about* some topic, and then characterizing the different patterns into which the collection of postings about the topic may fall. We propose that topics are mostly composed of a union of *chatter* (ongoing discussion whose subtopic flow is largely determined by decisions of the authors) and *spikes* (short-term, high-intensity discussion of real-world events that are relevant to the topic). We develop a generative model to capture this observed structure.

Individuals: Though the advent of personal publication gives everyone the same reach, not all individuals have the same grasp. We observe in our data that there are several distinct categories of individuals, as viewed by their impact on information diffusion through blogspace. This characterization allows us to predict the pattern of postings (about a topic) based on historical observations about the behavior of individuals in other contexts. We develop a model for propagation based on previous work in the area, and an algorithm for learning model parameters from observations. We apply this model to a large dataset, and report the findings.

2. RELATED WORK

The propagation of information has been studied extensively in the context of *gossiping* and *broadcasting* [18] in networks of a variety of forms, but the focus of that literature is essentially algorithmic in nature. Here, we are interested in models of information dispersion in which nodes in the network may or may not be interested in the information, and thus may or may not pass along the information to their neighbors.

The problem of understanding diffusion through a population has been studied in a number of communities, ranging from thermodynamics to epidemiology to marketing. Maxwell and others were the first to provide a rigorous analysis of this problem, in the context of thermodynamics. In that and subsequent work, statistical mechanics has looked at various models for the diffusion of particles of one gas in an other. Though this setting is superficially different, we find much that we can borrow from that field if we look at information as a kind of particle.

2.1 Information propagation and epidemics

Much previous research investigating the flow of information through networks has been based upon the observation of a deep analogy between the spread of disease and the spread of information in networks. The analogy between infection and information allows one to bring results of centuries of study of epidemiology to bear on questions of information diffusion. (See, for example, the book of Bailey [4] for some of the extensive work in this field.)

The classical disease-propagation models in epidemiology are based upon the cycle of disease in a host: a person u is first *susceptible (S)* to the disease, and, if u is then exposed to the disease by an infectious contact, then u herself becomes *infected (I)* (and *infectious*) with some probability p. The disease then runs its course in host u, and u is then *recovered (R)* (or *removed*, depending on the virulence of the disease). A recovered individual is *immune* to the disease for some period of time, but the immunity may eventually wear off, leaving u once again susceptible. Thus *SIR* models diseases in which recovered hosts are never again susceptible to the disease—as with a disease conferring lifetime immunity, like chicken pox, or a highly virulent disease from which the host does not recover—while *SIRS* models the situation in which a recovered host eventually becomes susceptible again—as with influenza, e.g. An important parameter of a network is its *epidemic threshold*: what is the minimum transmission probability ρ so that a disease spreads to infect a constant fraction of the network if a single seed node is initially infected? (In the model we consider in Section 5.2, unlike the typical model in epidemiology, the transmission probability $p = p(u, v)$ varies from edge to edge in the network.)

In blogspace, one might interpret the SIRS model as follows: initially, person u is not interested in topic x, but may become interested (S); u is actively interested in and posting on topic x (I); u has tired of topic x and is no longer posting on it (R); and u has forgotten her boredom, and now may potentially become interested in topic x again (S). For example, Girvan et al. [13] study a SIR model *with mutation*, in which a node u is immune to any strain of the disease which is sufficiently close to a strain with which u was previously infected. They observe that (with appropriate settings of parameters) it is possible to generate periodic outbreaks, where the disease oscillates between periods of epidemic outbreak and periods of calm while it mutates into a sufficiently new form that it can cause another major outbreak. In blogspace, one could imagine a blogger writing about Arnold *qua* movie star, growing bored of the topic, and then, after the topic of Arnold has evolved sufficiently, beginning to blog again about Arnold *qua* governor. (We observe this kind of ebb and flow in the popularity of various "spiky chatter"-type memes. See Section *4.2.1*.)

The majority of the epidemiology literature, including the work of Girvan et al. [13], focuses on the case of "fully mixed" or "homogeneous" networks, in which a node's contacts at any time step are chosen randomly from all other nodes in the population—i.e., there is no underlying network defining the contacts of each node. More recently, as the importance of network structure has become more clear, studies have begun to explore disease and information propagation on models of realistic networks.

In a model of small-world networks defined by Watts and Strogatz [31], Moore and Newman [24] calculate the epidemic threshold. However, this model does not account for some interesting and seemingly very important properties of real networks. A *power-law* network is one in which the probability that the degree of a node is k is proportional to $k^{-\alpha}$, for a constant α typically between 2 and 3. Power laws have been observed in many important real-world networks [23], including the social network defined by blog-to-blog links [22]. We now review some previous research on epidemic spreading on networks that follow a power law.

Pastor-Satorras and Vespignani [28] analyze an SIS model of (computer) virus propagation in power-law networks, showing that—in stark contrast to random or regular networks—the epidemic threshold is *zero*. (In other words, for any probability $\varepsilon > 0$ of disease transmission across an edge of the network, an epidemic will occur!) The epidemic threshold of power-law networks has also an interpretation in terms of the robustness of the network to random edge failure. Suppose that each edge in the network is deleted independently with probability $(1 - \varepsilon)$; we consider the network "robust" if most of the nodes are still connected. It is easy to see that nodes that remain in the same component as some initiator v_0 after the edge deletion process are exactly the same nodes that v_0 infects according to the disease transmission model above. This question has been considered from the perspective of *error tolerance* of networks like the Internet: what happens to the network if a random $(1 - \varepsilon)$-fraction of the links in the Internet fail? Many researchers have observed that power-law networks exhibit extremely high error tolerance [2; 7].

These results suggest that modeling information dispersion in blogs using this kind of transmission model is insufficient, since this falsely predicts that almost every node in the network will become "infected" with a topic if there is a non-zero probability of picking up a topic from a neighbor. One refinement is to consider a more accurate model of power-law networks. Eguíluz and Klemm [12] have demonstrated a non-zero epidemic threshold under the SIS model in power-law networks produced by a certain generative model that takes into account the high *clustering coefficient*—the proportion of triangles that are "closed," i.e., the probability that two people v and w will be friends if they have a common friend u—found in real social networks [31].

One can also resolve this discrepancy by a modification to the model of transmission. Wu et al. [33] consider the flow of information through real and synthetic email networks (generated according to a power-law distribution) under a model in which the probability that a node u will forward a meme to a neighbor v of u decays as the graph distance $d(s, u)$ from the original source node s of the meme increases. (The model is inspired by the observation of *homophily* in social networks: a person is biased towards having friends with similar interests to her own.) They observe that meme outbreaks under this model are typically limited in scope—unlike in the corresponding model without decay, where the epidemic threshold is zero—exactly as one observes in real data. Newman et al. [27] have also empirically examined the simulated spread of email viruses by examining the network defined by the email address books of a user

community.

2.2 The diffusion of innovation

The spread of a piece of information through a social network can also be viewed as the propagation of an *innovation* through the network. (For example, the URL of a website that provides an new, valuable service is such a piece of information.) Thus we can speak of bloggers *adopting* a topic t, analogous to adopting a new technology like, for example, blogs themselves.

In the field of sociology, there has been extensive study of the *diffusion of innovation* in social networks, examining the role of the process of *word of mouth* in spreading innovations. At a particular point in time, some nodes in the network have adopted the innovation, and others have not. Two fundamental models for the process by which new nodes adopt have been considered in the literature:

- *Threshold models [15].* Each node u in the network chooses a *threshold* $t_u \in [0, 1]$, typically drawn from some probability distribution. Every neighbor v of u has a nonnegative *connection weight* $w_{u,v}$ so that $\sum_{v \in \Gamma(u)} w_{u,v} \leq 1$, and u adopts if and only if $t_u \leq \sum_{\text{adopters } v \in \Gamma(u)} w_{u,v}$.
- *Cascade models [14].* Whenever a social contact $v \in \Gamma(u)$ of a node u adopts, then u adopts with some probability $p_{v,u}$. (In other words, every time a person close to a person u adopts, there is a chance that u will decide to "follow" v and adopt as well.)

In the *Independent Cascade model* of Goldenberg, Eitan, and Muller [14], we are given a set of N nodes, some of which have already adopted. At the initial state, some non-empty set of nodes are "activated." At each successive step, some (possibly empty) set of nodes become activated. The episode is considered to be over when no new activations occur. The set of nodes are connected in a directed graph with each edge (u, v) labeled with a probability $p_{u,v}$. When node u is activated in step t, each node v that has an arc (u, v) is activated with probability $p_{u,v}$. This influence is independent of the history of all other node activations. (If v is not activated in that time step, then u will never activate v.) The *General Cascade model* of Kempe, Kleinberg, and Tardos [19] generalizes the Independent Cascade model—and also simultaneously generalizes the threshold models described above—by discharging the independence assumption.

Kempe et al. are interested in a related problem on social networks with a marketing motivation: assuming that innovations propagate according to such a model, and given a number k, find the k "seed" nodes S_k^* that maximize the expected number of adopters of the innovation if S_k^* adopt initially. (One can then give free samples of a product to S_k^*, for example.)

3. CORPUS DETAILS

One of the challenges in any study involving tens of thousands of publishers is the tracking of individual publications. Fortunately for us, most of the publishers, including the major media sources, now provide descriptions of their publications using *RSS* (*rich site summary*, or, occasionally, *really simple syndication*) [20]. RSS, which was originally developed to support the personalization of the Netcenter portal, has now been adopted by the weblog community as a simple mechanism for syndication. In the present work, we focus on RSS because of its consistent presentation of dates—a key feature for this type of temporal tracking.

Our corpus was collected by daily crawls of 11,804 RSS blog feeds. We collected 2K–10K blog postings per day—Sundays were low,

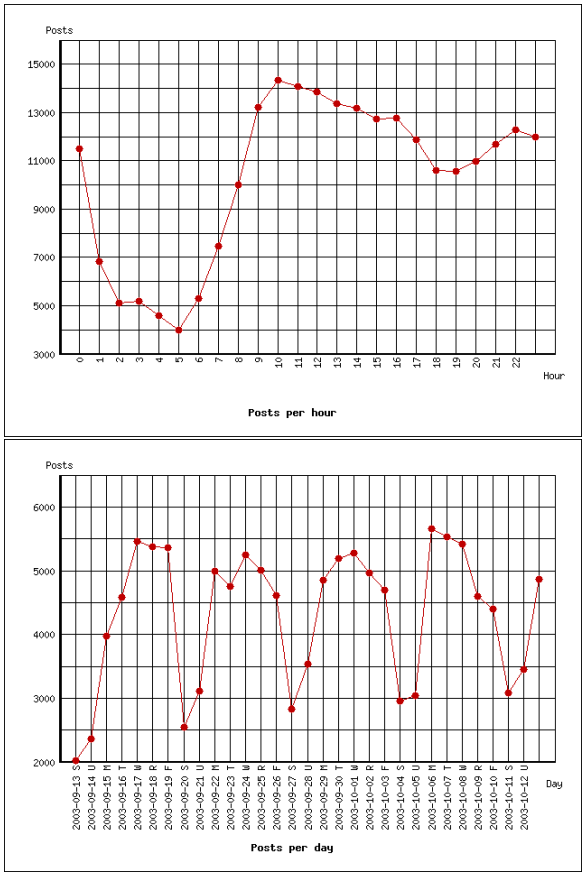

Figure 1: Number of blog postings (a) by time of day and (b) by day of week, normalized to the local time of the poster.

Wednesdays high—across these blogs, for a total of 401,021 postings in our data set. (Each posting corresponds to an "item" entry in RSS.) Complementing this, we also crawled 14 RSS channels from `rss.news.yahoo.com` hourly, to identify when topics were being driven by major media or real-world events, as opposed to arising within blogspace itself. The blog entries were stored as parent/child entities in WebFountain [32] and analyzed with a half-dozen special-purpose blog annotators to extract the various date formats popular in RSS, convert to UTF8, detag, etc.

See Figure 1 for the profile of blog postings within a day and from day-to-day, normalized by the poster's time zone. The most frequent posting is at 10AM. There is a pronounced dip at 6 and 7PM (the commute home? dinner? Must-See-TV?), an odd plateau between 2 and 3AM and a global minimum at 5AM. Posting seems to peak midweek, and dips considerably on weekends.

4. TOPIC CHARACTERIZATION AND MODELING

In this section, we explore the *topics* discussed in our data. We differentiate between two families of models: (i) *horizon* models, which aim to capture the long-term changes (over the course of months, years, or even decades) in the primary focus of discussion even as large chatter topics—like Iraq and Microsoft, as of this writing—wax and wane; and (ii) *snapshot* models, which focus on short-term behavior (weeks or months) while the background "chatter" topics are assumed to remain fixed. This paper explores snapshot models; we do not address horizon models, but instead

raise the issue as an interesting open problem.

4.1 Topic Identification and Tracking

To support our goal of characterizing topic activity, we must first find and track topics through our corpus. The field of *topic detection and tracking* has studied this problem in depth for a number of years—NIST has run a series of workshops and open evaluation challenges [30]; see also, for example, [3]. Our requirements are somewhat different from theirs; we require schemes that provide views into a number of important topics at different levels (very focused to very broad), but rather than either high precision or high recall, we instead require that our detected set contain good representatives of all classes of topics. We have thus evaluated a range of simple techniques, chosen the ones that were most effective given our goals, and then manually validated different subsets of this broader set for use in particular experiments.

Our evaluations of these different techniques revealed some unexpected gaps in our intuition regarding blogspace; we give a brief walkthrough here. First, we treated references to particular websites as topics, in the sense that bloggers would read about these "interesting" sites in another blog and then choose to write about them. However, while there are over 100K distinct links in our corpus, under 700 of them appear 10 times or more—not enough to chart statistically significant information flows. Next, we considered recurring sequences of words using sequential pattern mining [1]. We discovered under 500 such recurrent sequences, many of which represented automatically generated server text, or common phrases such as "I don't think I will" and "I don't understand why." We then turned to references to entities defined in the TAP ontology [16]. This provided around 50K instances of references to 3700 distinct entities, but fewer than 700 of these entities occurred more than 10 times. The next two broader sets provided us with most of the fodder for our experiments. We began with a naive formulation of proper nouns: all repeated sequences of uppercase words surrounded by lowercase text. This provided us with 11K such features, of which more than half occurred at least 10 times. Finally, we considered individual terms under a ranking designed to discover "interesting" terms. We rank a term t by the ratio of the number of times that t is mentioned on a particular day i (the term frequency $tf(i)$) to the average number of times t was mentioned on previous days (the cumulative inverse document frequency). More formally, $tfcidf(i) = (i-1)tf(i)/\sum_{j=0}^{i-1} tf(j)$. Using a threshold of $tf(i) > 10$ and $tfcidf(i) > 3$ we generate roughly 20,000 relevant terms.

All features extracted using any of these methods are then spotted wherever they occur in the corpus, and extracted with metadata indicating the date and blog of occurrence.

4.2 Characterization of Topic Structure

To understand the structure and composition of topics, we manually studied the daily frequency pattern of postings containing a large number of particular phrases. We analyzed the 12K individual words most highly ranked under the tfcidf ranking described above. Most of these graphs do not represent topics in a classical sense, but many do. We hand-identified 340 classical topics, a sample of which is shown in Table 1.

Next, based on our observations, we attempt to understand the structure and dynamics of topics by decomposing them along two orthogonal axes: *chatter*, internally driven, sustained discussion; and *spikes*, externally induced sharp rises in postings. We then refine our model by exploring the decomposition of these spikes into subtopics, so that a topic can be seen as the union of chatter and spikes about a variety of subtopics.

apple	arianna	ashcroft	astronaut
blair	boykin	bustamante	chibi
china	davis	diana	farfarello
guantanamo	harvard	kazaa	longhorn
schwarzenegger	udell	siegfried	wildfires
zidane	gizmodo	microsoft	saddam

Table 1: Example topics identified during manual scan.

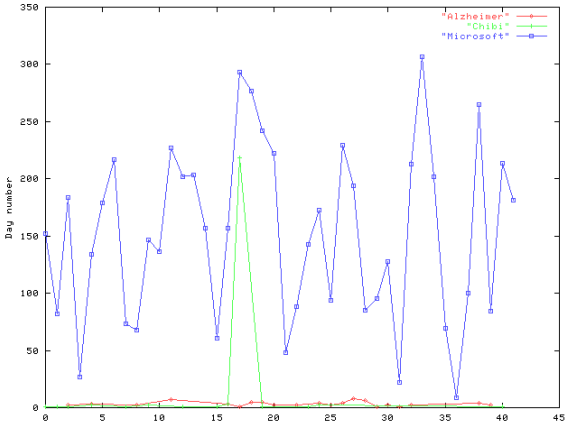

Figure 2: Three types of topic patterns: the topic "Chibi" (green) is *Just Spike*; "Microsoft" (blue) is *Spiky Chatter*; and "Alzheimer's" (red) is *Mostly Chatter*.

4.2.1 *Topic = Chatter + Spikes*

There is a community of bloggers interested in any topic that appears in postings. On any given day, some of the bloggers express new thoughts on the topic, or react to topical postings by other bloggers. This constitutes the *chatter* on that topic.

Occasionally, an event occurring in the real world induces a reaction from bloggers, and we see a *spike* in the number of postings on a topic. Spikes do not typically propagate through blogspace, in the sense that bloggers typically learn about spikes not from other blogs, but instead from a broad range of channels including mainstream media. Thus, we can assume all informed authors are aware of the topical event and have an opportunity to write about it.

On rare occasions, the chatter reaches *resonance*, i.e., someone makes a posting to which everyone reacts sharply, thereby causing a spike. The main characteristic of resonance is that a spike arises from either no external input or a very small external input. The formation of order (a spike) out of chaos (chatter) has been observed in a variety of situations [29], though observation of our data reveals that this happens very rarely in blogspace. In fact, the only sustained block re-posting meme that we observed in our data consisted of the "aoccdrnig to rscheearch at an elingsh uinervtisy it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer is at the rghit pclae" story which came out of nowhere, spiked and died in about 2 weeks (with most postings over a four-day period).

Depending on the average chatter level and pertinence of the topic to the real world, topics can be roughly placed into one of the following three categories, with examples shown in Figure 2:

Just Spike: Topics which at some point during our collection window went from inactive to very active, then back to inactive. These topics have a very low chatter level. E.g., Chibi.

windows	server	services	longhorn
exchange	ie	office	msdn
outlook	msn	gates	redmond
eolas	xp	netscape	powerpoint
scoble	pdc	motorola	avalon
ms	vb	acrobat	xaml

Table 2: Top coverage terms for Microsoft spikes.

Spiky Chatter: Topics which have a significant chatter level and which are very sensitive to external world events. They react quickly and strongly to external events, and therefore have many spikes. E.g., Microsoft.

Mostly Chatter: Topics which were continuously discussed at relatively moderate levels through the entire period of our discussion window, with small variation from the mean. E.g., Alzheimer's.

Spiky Chatter topics typically have a fairly high level of chatter, with the community responding to external world events with a spike; their persistent existence is what differentiates Spiky Chatter from spikes. They consist of a superposition of multiple spikes, plus a set of background discussion unrelated to any particular current event. For example, the Microsoft topic contains numerous spikes (for example, a spike towards the end of our window around a major announcement about Longhorn, a forthcoming version of Windows) plus ongoing chatter of people expressing opinions or offering diatribes regarding the company and its products.

4.2.2 *Topic = Chatter + Spiky Subtopics*

In this section, we refine our model of Topic = Chatter + Spikes by examining whether the spikes themselves are decomposable. Intuitively, the community associated with a topic can be seen as randomly choosing a subtopic and posting about it. When an external world event occurs, it is often particular to something very specific—that is, a subtopic—especially for complex topics. In this section, we consider a subtopic-based analysis using the spikes in the complex, highly posted topic "Microsoft" as a case study. Microsoft was especially appropriate for this analysis, as several Microsoft-related events occurred during the collection of our data set, including the announcement of blog support in Longhorn.

We used a multi-step process to identify some key terms for this experiment. First, we looked at every proper noun x that co-occurred with the target term "Microsoft" in the data. For each we compute the support s (the number of times that x co-occurred with the target) and the reverse confidence $c_r := P(target|x)$.

Thresholds for s and c_r were manipulated to generate rational term sets. As is common with these cases, we do not have a hard-and-fast support and confidence algorithm, but found that s in the range of 10 to 20 and c_r in the range of 0.10 to 0.25 worked well. For the target "Microsoft," this generates the terms found in Table 2. Of course, this is not a complete list of relevant subtopics, but serves rather as a test set. For these terms, we looked at their occurrences, and defined a spike as an area where the posts in a given day exceeded $\mu + 2\sigma$. We then extended the area to either side until a local minimum less than the mean was reached. We refer to posts during these intervals as *spike posts*.

Now, having identified the top coverage terms, we deleted spike posts related to one of the identified terms from the Microsoft topic. The results are plotted in Figure 3. The de-spiked posts line shows a considerable reduction in the spikes of the Microsoft graph, with minor reduction elsewhere. Note that even in the spiky area we

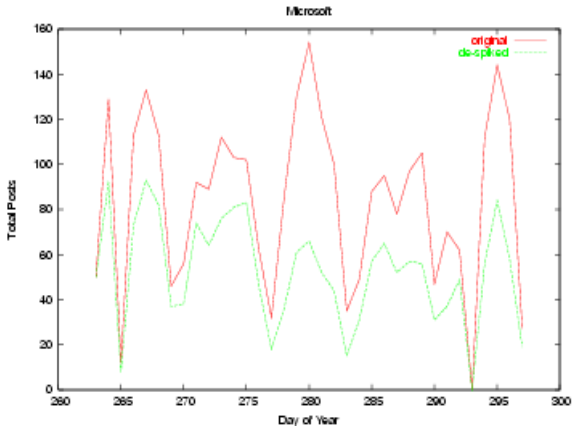

Figure 3: The topic density for posts on Microsoft, both before and after spike removal.

series	server	os	longhorn
pc	ie	mac	gui
apple	jobs	dell	ui
ram	xp	explorer	drm
unix	pcs	linux	apples
ms	macs	quicktime	macintosh

Table 3: Top coverage spike terms for Windows. Terms on a grey background are also spike terms for Microsoft (Table 2).

are not getting a complete reduction, suggesting we may not have found all the synonymous terms for those spike events, or that subtopic spikes may be correlated with a latent general topic spike as well.

This analysis in no way implies that the topics in Table 2 are atomic. We also explored the subtopic "Windows"—one of the subtopics with better coverage—and looked at its decomposition. The proper noun selection was performed as before, generating the term set in Table 3. There is some duplication of terms from Table 2, as the topics "Microsoft" and "Windows" overlap significantly. However, some terms unique to Windows appear, especially the comparison to Apple (Apple, Steve Jobs, Quicktime, Mac, Macs, Macintosh). Applying these terms to the Windows posting frequency, we see the results in Figure 4. Again, we see a similar reduction in spikes, indicating that we have found much of the spiky behavior of this topic. As might be expected with a more focused topic, the top 24 spike terms have better coverage for "Windows" than for "Microsoft," leaving a fairly uniform chatter.

This case study strongly supports our notion of a spike and chatter model of blog posting. While not presented here, similar behavior was observed in a number of other topics (terrorism, Linux, the California recall election, etc.).

4.2.3 *Characterization of Spikes*

Having presented a qualitative decomposition of topics into chatter and spikes, we now present measurements to quantify the nature of these spikes. Each chatter topic can be characterized by two parameters corresponding to the chatter level (distribution of the number of posts per day) and the spike pattern (distribution of the frequency, volume, and shape of spikes).

To perform these evaluations, we hand-tagged a large number of topics into the categories given in Section *4.2.1*. Of those hand-

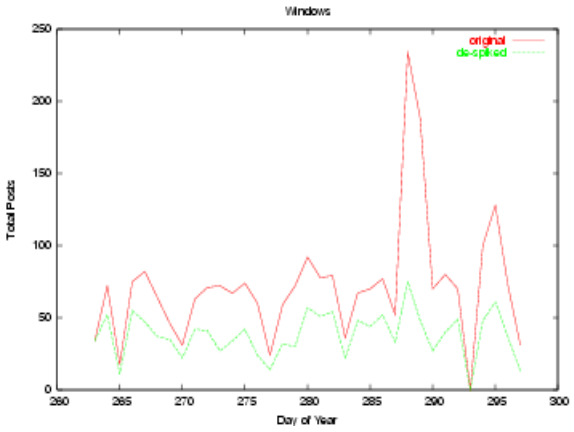

Figure 4: The topic density for posts on Windows, both before and after spike removal.

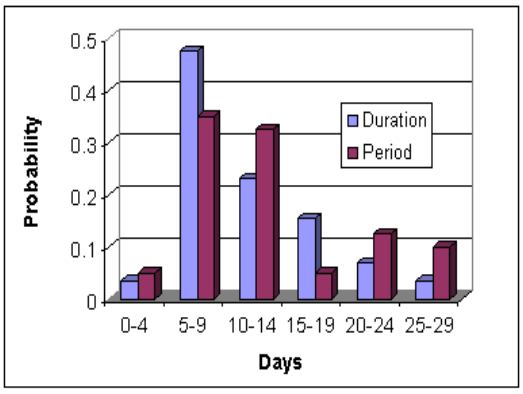

Figure 5: Distribution of spike duration and period within chatter topics.

tagged topics, 118 fell into the chatter category; we performed this characterization study on those topics. We used the simple spike definition of Section *4.2.2* to determine where the spikes occurred in each chatter topic; an examination of the spikes found by this algorithm led us to believe that, while simple, it indeed captures our intuition for the spikes in the graph.

To begin, the average number of posts per day for non-spike regions of our collection of chatter topics ranges between 1.6 to 106. The distribution of non-spike daily average is well-approximated by $\Pr[\text{average number of posts per day} > x] \sim ce^{-x}$.

Next, we focus on characteristics of spike activity. Figure 5 shows the distribution of spike durations and periods. Most spikes in our hand-labeled chatter topics last about 5–10 days. The median period between spike centers is about two weeks.

Figure 6 shows the distribution of average daily volume for spike periods. The median spike among our chatter topic peaks at 2.7 times the mean, and rises and falls with an average change of 2.14 in daily volume.

5. CHARACTERIZATION AND MODELING OF INDIVIDUALS

We have covered the high-level statistical "thermodynamic" view of the data in terms of aggregates of posts at the topic level; now

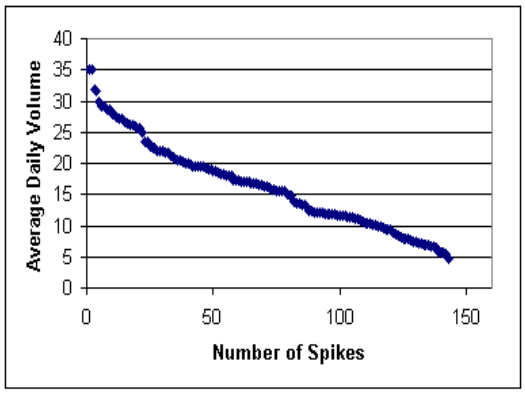

Figure 6: Average daily volume of spikes within chatter topics.

Region	Fraction of topics
RampUp	3.7%
RampDown	5.1%
MidHigh	9.4%
Spike	18.2%

Table 4: Fraction of topics containing each region type.

we turn to a view more akin to particle dynamics, in which we attempt to uncover the path of particular topics through the various *individuals* who make up blogspace. We begin in Section 5.1 by characterizing individuals into a small number of classes, just as we did for topics in the previous section. Next, in Section 5.2 we formulate a model for propagation of topics from person to person through blogspace, and we present and validate an algorithm for inducing the model. Finally, we apply the model to real data, and give some preliminary applications.

Our model is akin to traditional models of disease propagation, in which individuals become "infected" by a topic, and may then pass that topic along to others with whom they have close contact. In our arena, close contact is a directed concept, since a may read the blog of b, but not vice versa. Such a model gives a thorough understanding of how topics may travel from person to person. Unfortunately, we do not have access to direct information about the source that inspired an author to post a message. Instead, we have access only to the surface form of the information: the sequence in which hundreds, thousands, or tens of thousands of topics spread across blogspace. Our algorithm processes these sequences and extracts the most likely communication channels to explain the propagation, based on the underlying model.

5.1 Characterizing Individuals

We begin with a quick sense of the textual output of our users. Figure 7 shows the distribution of the number of posts per user for the duration of our data-collection window. The distribution closely approximates the expected power law [23].

We now wish to classify these users. We adopt a simple set of predicates on topics that will allow us to associate particular posts with parts of the lifecycle of the topic. Given this information, we will ask whether particular individuals are correlated with each section of the lifecycle. The predicates are defined in the context of a particular time window, so a topic observed during a different time window might trigger different predicates. See Table 6 for the definitions of these predicates.

Table 4 shows the fraction of topics that evince each of these re-

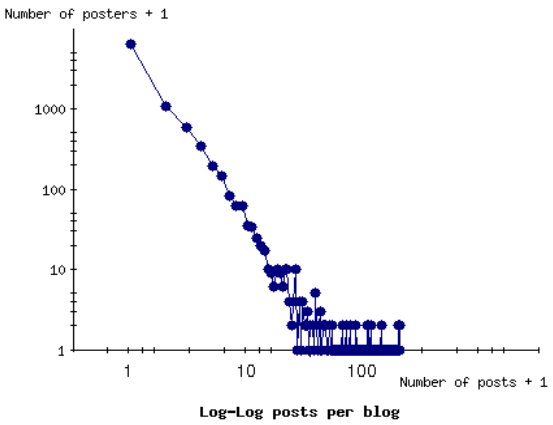

Figure 7: Distribution of number of posts by user.

Region	Up	Down	Mid	Spike
Users with > 4 posts and $> \mu + 3\sigma$	20	55	157	310
Total posts this region	1733	3300	12453	55624

Table 5: Number of users associated with each region.

gions. We can then attempt to locate users whose posts tend to appear in RampUp, RampDown, MidHigh, or Spike regions of topics. However, we must exercise caution in tracking this correspondence: for example, we wish to avoid capturing users who simply happened to post more frequently during the early part of our data-collection window, and thus are more likely to post during regions identified as RampUp by our predicates. To overcome this difficulty, we consider the probability p_i that a post on day i falls into a given category (e.g., RampUp). For any given user, we then consider the pair (t_i, c_i) of total posts on day i and posts in the category on day i, respectively. The total number of posts in the category is $C = \sum_i c_i$. We can then define a "random" user who contributes the same number of posts each day, but does so without bias for or against the category. The expected number of posts in the category for the random user is then $\sum_i p_i t_i$. Because the random user produces a sum of independent random variables, each of which is simply a series of Bernoulli trials with some bias depending on the day, we can determine the probability that the random user would produce C or more posts in the category, and therefore determine the extent to which we should be surprised by the behavior of the given user. We set our threshold for surprise when the number of occurrences is more than three standard deviations beyond the mean of the random user.

Using this technique, we give the number of users who are unusually strong contributors to each region in Table 5. In some cases, as for the Up region, the numbers are relatively low, but the total number of posts in the region is also quite small. The correlation is quite strong, leading us to suggest that evaluating broader definitions of a "ramp up" phase in the discussion of a topic may identify a larger set of users correlated with this region. For regions such as Mid or Spike, the number of associated users is quite substantial, indicating that there are significant differing roles played by individuals in the lifecycle of a topic.

5.2 Model of Individual Propagation

We derive our formal model from the Independent Cascade model of Goldenberg et al. [14] and generalized to the General Cascade

Predicate	Algorithm	Region
RampUp	All days in first 20% of post mass below mean, and average day during this period below $\mu - \sigma/2$.	First 20% of post mass.
RampDown	All days in last 20% of post mass below mean, and average day during this period below $\mu - \sigma/2$.	Last 20% of post mass.
MidHigh	All days during middle 25% of post mass above mean, and average day during this period above $\mu + \sigma/2$.	Middle 25% of post mass.
Spike	For some day, number of posts exceeds $\mu + 2\sigma$.	From spike to inflection point below μ, both directions.

Table 6: Lifecycle predicates on topics.

Model by Kempe et al. [19]. We are given a set of N nodes, corresponding to the authors. At the initial state of each episode, some (possibly empty) set of nodes have written about the topic. At each successive state, some (possibly empty) set of authors (including possibly some who have already written before) write about the topic. The episode is considered to be over when no new articles appear for some number of time steps, the *Timeout Interval*.

With the Independent Cascade Model, the set of authors are connected in a directed graph with each edge labeled with a probability. When author v writes an article at time t, each node w that has an arc from v to it writes an article about the topic with the probability $\kappa(v, w)$, the *copy probability*. This influence is independent of history whether any other neighbors of w have written an article. The General Cascade Model can be seen as generalizing this by eliminating the assumption of independence.

We introduce the notion that a user may visit certain blogs frequently, and other blogs infrequently; we capture this with an edge property $r_{u,v}$, denoting the probability that u reads v on any given day. We also introduce the notion of *stickiness* of a topic, S—more sticky topics are more likely to infect the reader.[1]

Formally, propagation in our model occurs as follows. If a topic exists at vertex u on a given day, then we compute the probability that it will propagate from u to a neighboring vertex v as follows. Node v reads the topic from node u on any given day with reading probability $r_{u,v}$, so we choose a delay from an exponential distribution with parameter $r_{u,v}$. With probability S, the stickiness of the topic, the topic will "stick" with v. And finally, with probability $\kappa_{u,v}$, the *copy probability*, the author of v will choose to write about it. If v reads the topic and it does not stick, or is not copied, then v will never choose to copy that topic from u; there is a single opportunity for the topic to propagate down any given edge.

Alternately, one may imagine that once u is infected, v will become infected with probability $S\kappa_{u,v}r_{u,v}$ on any given day, but once the $r_{u,v}$ coin comes up heads, no further trials are made.

Thus, given the transmission graph (and, in particular, the reading frequency r and the copy probability κ for each edge), and given the stickiness S of a particular meme, the distribution of propagation patterns is now fully established. Given a community and a timeout interval, our goal is therefore to learn the arcs and associ-

[1] Stickiness of a topic is analogous to *virulence* in the disease propagation literature.

ated probabilities from a set of episodes. Using these probabilities, given a new episode, we would like to estimate the stickiness of the new episode from an initial fragment of the episode. Then, we would like to be able to predict the propagation pattern that will be associated with the episode.
We now present a few possible extensions to the model:

- Most topics do not travel exclusively through blogspace; rather, they are real-world events that are covered to some extent in the media. During online coverage of the topic, certain bloggers may read about the topic in other blogs and respond, while others may read about the topic in the newspaper and write without reference to other weblogs. Our model can be extended by introducing a node corresponding to the "real world" which we view as writing about a topic whenever the topic is covered sufficiently in the media. Transmission probabilities and delays are handled as they are elsewhere in the model, but it is assumed that essentially all bloggers may receive input from this "real world" node.

- In the real world, communities can become quite large, and most people do not have the time to read more than a few blogs on any regular basis. This phenomenon can be modeled either by limiting the indegree of nodes, or by allowing only some small number of in-edges to influence a particular node at any time step. The model can be extended by adding an additional *Attention Threshold* (AT) parameter.

  More sophisticated models can capture the fact that the attention threshold is a function of the other episodes (in the same or other communities) that are occurring at the same time—the more concurrent episodes, the lower the attention threshold for each episode. This can explain the phenomenon that during high-chatter events such as the Iraq war or the California elections, many other topics that would otherwise have received a lot of attention in fact received little.

5.3 Induction of the Transmission Graph

In the following, we make a *closed world assumption* that all occurrences of a topic other than the first one are the result of communication via edges in the model. As described above, this assumption can be weakened by introducing an "outside world" node with appropriate parameters into the model.
A *topic* in the following is a URL, phrase, name, or any other representation of a meme that can be tracked from page to page. We gather all blog entries that contain a particular topic into a list $[(u_1, t_1), (u_2, t_2), \ldots, (u_k, t_k)]$ sorted by publication date of the blog, where u_i is the universal identifier for blog i, and t_i is the time at which blog u_i contained a reference to the topic. We refer to this list as the *traversal sequence* for the topic.
We shall make critical use of the following observation: we wish to induce the relevant edges among a candidate set of $\Theta(n^2)$ edges, and we have only limited data, but the fact that blog a appears in a traversal sequence, and blog b *does not* appear later in the same sequence gives us evidence about the (a, b) edge—that is, if b were a regular reader of a's blog with a reasonable copy probability, then sometimes memes discussed by a should appear in b's blog. Thus, we gain information from both the presence and absence of entries in the traversal sequence.
We present an iterative algorithm to induce the transmission graph. Assume that we have an initial guess at the value of (r, κ) for each edge, and we wish to improve our estimate of these values. We adopt a two-stage process:

Step 1: Using the current version of the transmission graph, compute for each topic and each pair (u, v) the probability that the topic traversed the (u, v) edge.

Step 2: For fixed u and v, recompute (r, κ) based on the posterior probabilities computed above.

5.3.1 Step 1

We are given as input the traversal sequence for a particular topic. For each v in the sequence, we consider all previous vertices u in the sequence, and compute $\Pr(u \to v)$, the probability that the topic would have traversed from u to v given the delay between u and v in the sequence. We then normalize by the sum of these probabilities to compute posteriors over all nodes u of the probability that each was v's source of inspiration. That is, setting $r = r_{u,v}$, $\kappa = \kappa_{u,v}$, and δ to be the delay in days between u and v:

$$p_{u,v} = \frac{r(1-r)^{\delta}\kappa}{\sum_{w<v} r_{w,v}(1-r_{w,v})^{\delta_{w,v}}\kappa_{w,v}}$$

In practice, for efficiency reasons, we consider only the 20 values of w closest to v, and require propagation to occur within 30 days.

5.3.2 Step 2

We perform the following operation for each fixed u, v.
First, we require a sequence S_1 of triples (p, δ, s), each corresponding to some topic appearing in u and then v, where p is the posterior probability that the topic traveled from u to v as computed above, δ is the delay in days between the appearance of the topic in u and in v, and s is the stickiness of the topic. We also require a sequence S_2 of pairs (Δ, s) for topics with stickiness s in which u appeared, v did not appear later in the sequence, and Δ days elapsed between the appearance of u and the end of our snapshot.
We can then estimate an updated version of r, κ as follows:

$$r = \frac{\sum_i p_i}{\sum_i p_i \delta_i}$$

$$\kappa = \frac{\sum_i p_i}{\sum_{i \in S_1} \Pr[r \le \delta_i] + \sum_{i \in S_2} \Pr[r \le \Delta_i]}$$

where $\Pr[a \le b] = (1-a)(1-(1-a)^b)$ is the probability that a geometric distribution with parameter a has value $\le b$.

5.3.3 Iteration and Convergence

We now have an improved guess at the transmission graph, so we can return to step 1 and recompute posteriors, cycling through the process until convergence. During step 1, we use our model of the graph to guess how data traveled. During step 2, we use our guess about how data traveled to improve our model of the graph.
For our data sets, the values of r and κ converge within between 2 and 5 iterations, depending on the data, to a vector of values within 1% of the limiting value under the L_2 norm.

5.3.4 Synthetic Validation of the Algorithm

In order to validate the algorithm, we created a synthetic series of propagation networks, ran each synthetic network to generate observable sequences of infection by particular topics, and then ran our mining algorithm to extract back the underlying propagation network. The synthetic graphs are modified Erdös-Renyi random graphs:[2] a number of vertices n is fixed, as is a target degree d.

[2]In the full version of this paper, we will also present synthetic benchmarks based on power law random graphs [6; 21].

Topics per node	μ_r	σ_r	μ_κ	σ_κ
2	0.718	0.175	0.141	0.455
4	0.703	0.157	0.107	0.039
6	0.694	0.134	0.103	0.034

Table 7: Mean and standard deviation for r and κ in low-traffic synthetic benchmark. Correct values: $\mu = 0.66, \sigma = 0.1$.

Each vertex selects d out-neighbors uniformly with replacement from the vertex set; all parallel edges and self-loops are then removed. Each edge is then given a (r, κ) value; we used $r = 2/3$ and $\kappa = 1/10$ for our tests.

We began with a series of graphs with $n = 1000$ and $d = 3$. For such graphs, we seeded a number of topics at each vertex, ranging from 20 to 60. Due to the small value of κ, we saw on average between 2 and 6 topics originating from each vertex. We considered only edges that were traversed by at least three topics with probability at least 0.1. We then compared the resulting edge set against the edge set from the original propagation network. An edge was counted as erroneous if it appeared in only one of those two graphs—in other words, in this benchmark we penalize for both missing edges and unnecessary edges. Of 3000 edges, the algorithm requires little data to infer the correct edges: once it saw 6 topics per node on average, it correctly inferred 2663 of the 3000 edges, plus 4 erroneous additional edges. For this benchmark, the algorithm converges in two iterations. The mean and standard deviation of the inferred values of r and κ for this experiment are shown in Table 7.

Next, we turn to a propagation model with higher degrees in which topics tend to take off and propagate throughout the graph, making it more difficult to learn exactly how the information had traveled. The parameters are $n = 500$, $d = 9$, and we take 20 topics per node. Topic sizes range from 1 to slightly over 200. The estimated r values have mean 0.73 and standard deviation 0.12; the κ values have mean 0.08 and standard deviation 0.03. The system identifies almost all relevant edges (to within 1%), and identifies a further almost 9% spurious edges due to the more complex structure of this task. Thus, both the edges and the estimated parameters of the edges are very close to the underlying model.

5.4 Validation and Analysis of Learned Parameters

Now that we have validated the algorithm on synthetic data, we validate the model itself against our data. We run the graph induction algorithm as described above on all the ProperName sequences in our dataset. As we have seen, roughly 20% of these sequences contain spikes, and fewer than 10% contain RampUp and RampDown areas. So the dataset consists of both signal and noise. Rather than introducing a "real world" node to modeling communication through the general media, we restrict our attention to topics for which at least 90% of the occurrences are in blogspace, rather than in our RSS media content. This focuses on about 7K topics.

To validate that the model has in fact discovered the correct edges, we performed two experiments. First, we downloaded the top 100 blogs as reported by `http://blogstreet.com`. Of the 100 blogs, 70 of them were in our RSS-generated dataset. We then used the model to rank individual nodes of the network based on the amount of traffic flowing through those nodes. Of the 70 nodes in our dataset, 49 were in the top 10% of blogs in our analysis; 40 were in the top 5%, and 24 were in the top 1.2%.

As a second validation, we ranked all edges in the final model by the expected number of topics that flowed down the edge, and produced the top 200. We hand-examined a random sample of this set, and in 90% of the cases were able to find a link between the two blogs. Note that we were able to make use of the structure of blogspace in the discovery of these links (i.e., blogrolls, and userids appearing inline), while the algorithm did not have access to these mechanisms, and made its determinations based on topics alone.

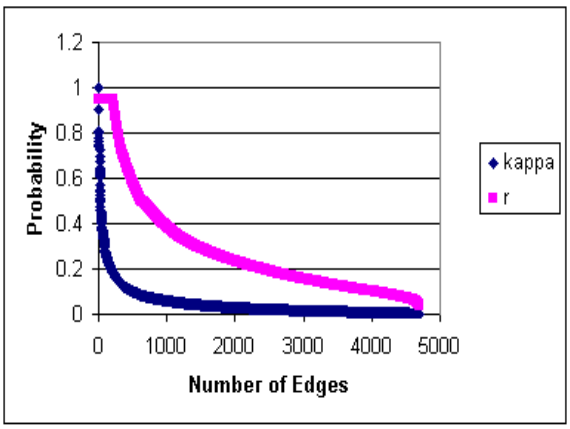

Figure 8: Distribution of Inverse Mean Propagation Delay (r) and Copy Probability (κ).

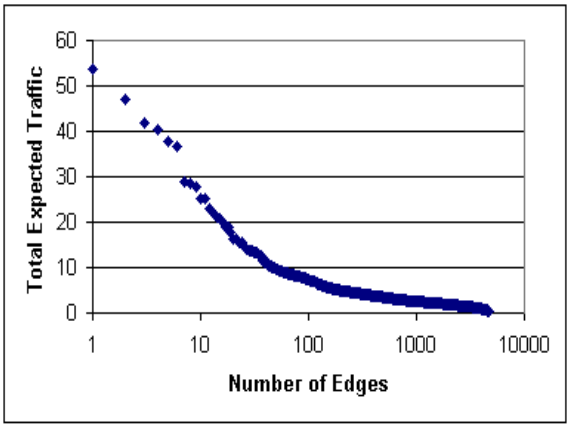

Figure 9: Expected Traffic.

5.4.1 *Parameters Learned by the Algorithm*

Figure 8 shows the distributions of r and κ as learned by the algorithm on the approximately 7K topics described above. Most edges have an expected propagation delay ($1/r$) of fewer than 5 days; the mean is 0.28 and the standard deviation is 0.22. Copy probabilities are quite low, with mean 0.04 and standard deviation 0.07, indicating that even bloggers who commonly read from another source are selective in the topics they choose to write about.

Figure 9 shows the distribution of expected traffic along each edge; i.e., over the set of 11K given topics, for a particular edge (a, b), how many times does b read about something on a and consequently write about it? The iteration converges to about 4000 edges with traffic. Popular edges might have 50 expected copies; the median edge has 1–2 total expected messages that traverse it.

6. CONCLUSIONS

Blogspace, by virtue of its fine grained observability, offers a fertile

testbed for developing and testing models of information diffusion, especially through the medium of personal publishing. In this paper, we showed how by using macro (topical) and micro (individual) models, various structures and behaviors can be understood, ranging from the strong driving effect of outside world events on what is being discussed to the applicability of traditional sociological models of influence to bloggers. Employing such characterizations allows applications to take advantage of these rapidly emerging web phenomena.

7. REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proc. ICDE*, pages 3–14, 1995.

[2] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, July 2000.

[3] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer, 2002.

[4] Norman Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 2nd edition, 1975.

[5] Venkatesh Bala and Sanjeev Goyal. A strategic analysis of network reliability. *Review of Economic Design*, 5:205–228, 2000.

[6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[7] Béla Bollabas and Oliver Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35, 2003.

[8] Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85:5468–5471, 2000. `cond-mat/0007300`.

[9] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628, November 2000. `cond-mat/0007048`.

[10] Paolo Crucitti, Vito Latora, Massimo Marchiori, and Andrea Rapisarda. Efficiency of scale-free networks: Error and attack tolerance. *Physica A*, 320:622–642, 2003.

[11] Jared Diamond. *Guns, Germs, and Steel*. Random House, 1997.

[12] Víctor M. Eguíluz and Konstantin Klemm. Epidemic threshold in structured scale-free networks. *Physical Review Letters*, 89(108701), 2002. `cond-mat/0205439`.

[13] Michelle Girvan, Duncan S. Callaway, M. E. J. Newman, and Steven H. Strogatz. A simple model of epidemics with pathogen mutation. *Phys. Rev. E*, 65(031915), 2002. `nlin.CD/0105044`.

[14] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[15] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1987.

[16] R. V. Guha and Rob McCool. TAP: A system for integrating web services into a global knowledge base.

[17] Hans Haller and Sudipta Sarangi. Nash networks with heterogeneous agents. Working Paper Series E-2001-1, Virginia Tech, 2003.

[18] Sandra M. Hedetniemi, Stephen T. Hedetniemi, and Arthur L. Liestman. A survey of gossiping and broadcasting in communication networks. *Networks*, 18:319–349, 1988.

[19] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proc. KDD*, 2003.

[20] Andrew King. The evolution of RSS. `http://www.webreference.com/authoring/languages/xml/rss/1/`.

[21] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. FOCS*, 2000.

[22] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *Proc. WWW*, pages 568–576, 2003.

[23] M. Mitzenmacher. A brief history of lognormal and power law distributions. In *Allerton Commun. Control Comput.*, 2001.

[24] Cristopher Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61:5678–5682, 2000. `cond-mat/9911492`.

[25] Stephen Morris. Contagion. *Review of Economic Studies*, 67, 2000.

[26] M. E. J. Newman. The spread of epidemic disease on networks. *Phys. Rev. E*, 66(016128), 2002. `cond-mat/0205009`.

[27] M. E. J. Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66(035101), 2002.

[28] Romauldo Pasto-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Letters*, 86(14):3200–3203, April 2001.

[29] Steven Strogatz. *Sync: The emerging science of spontaneous order*. Hyperion, 2003.

[30] Topic Detection and Tracking (TDT-2003). `http://www.nist.gov/TDT`.

[31] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[32] WebFountain. `http://www.almaden.ibm.com/WebFountain/`.

[33] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. Information flow in social groups. manuscript, 2003.

[34] H. Peyton Young. The diffusion of innovation in social networks. Sante Fe Institute Working Paper 02-04-018, 2002.