Information Spreading in Context

Dashun Wang[1,2] Zhen Wen[3] Hanghang Tong[3] Ching-Yung Lin[3]
Chaoming Song[1,2] Albert-László Barabási[1,2,4]
[1]CCNR, Dept. of Physics and Computer Science, Northeastern University, Boston, MA 02115, USA
[2]CCSB, Dana-Farber Cancer Institute, Harvard University, Boston, MA 02115, USA
[3]IBM T.J. Watson Research Center, Hawthorne, NY 10532, USA
[4]Dept. of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
{dashunwang,chaomingsong, barabasi}@gmail.com
{zhenwen,htong,chingyung}@us.ibm.com

ABSTRACT

Information spreading processes are central to human interactions. Despite recent studies in online domains, little is known about factors that could affect the dissemination of a single piece of information. In this paper, we address this challenge by combining two related but distinct datasets, collected from a large scale privacy-preserving distributed social sensor system. We find that the social and organizational context significantly impacts to whom and how fast people forward information. Yet the structures within spreading processes can be well captured by a simple stochastic branching model, indicating surprising independence of context. Our results build the foundation of future predictive models of information flow and provide significant insights towards design of communication platforms.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Measurement, Experimentation, Human Factors

Keywords

Information spread, Context, Social networks, Structure, Network science

1. INTRODUCTION

Information spreading plays an essential role in numerous human interactions, including the spread of innovations [32, 30], knowledge and information security management [15], social influence in marketing [10, 21, 25], and more. Thanks to the increasing availability of large-scale data, we have witnessed great advances in understanding how information propagates from person to person, ranging from incentivized word-of-mouth effects when recommending products [25, 19], to understanding how a single piece of information forms internet chain letters on a global scale [27].

Despite recent studies in online social networks [25, 2, 16, 26], it has been difficult to obtain detailed traces of information dissemination alongside relevant contextual data such as people's real social connections, their behavioral profiles, and job roles in organizations. Therefore, an important question is largely unanswered: *to what extent do spreading processes depend on the underlying social network and behavioral profiles of individuals*. Indeed, on one hand, information such as rumors, innovations and opinions diffuses through the underlying social networks. To whom and to how many people a user would pass such information is constrained by whom s/he connects to and how well s/he is connected in the social network, and the strength of those connections. On the other hand, the population-based heterogeneity in personal profiles coexists with complex connectivities between individuals, raising questions about to what degree the diverse profiles of individuals, from personal interests and expertise to communities and hierarchy, impact the information spreading process. Understanding the role of these features is of fundamental importance.

The lack of contextual information could change drastically, however, thanks to the pervasive use of email communications in well-documented settings, such as corporate work forces [11, 28, 12, 38, 4, 1, 22, 20]. Indeed, emails have become the most important communication method in various settings [36, 28], unveiling detailed traces of social interactions among large populations. Previous studies [36, 6] have shown that email communications serve as a good indicator of social ties. *Forwarded emails* [29], written by someone other than the sender and sent to someone who was not included in the original email, serve as an ideal proxy for the information spreading process, where the single piece of information, the original body of the email, is passed through the social network.

We compiled a new dataset by integrating two related but distinct data structures, collected from a large-scale, privacy-preserving distributed social sensor system. First, we collected two years of email communication data from 8,952 volunteer employees within a large technology firm operating in more than 70 countries. Emails occupy the majority of information workers' time and thus provide high-quality observation of the social context, i.e., the real social connections of employees in the workplace [36, 6]. In addition to such "informal networks," we investigated the "formal networks," imposed by the corporation such as their hierarchical structure, as well as demographic data such as geography, job role, self-specified interests, performance, etc. This dataset provides us unique opportunities to study the interplay between the information spreading and its context. This issue is largely not addressed in previous studies par-

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

tially due to the lack of such a multi-faceted dataset and the difficulty in matching user IDs across multiple sources.

Specifically, we investigate the impact of context on spreading processes in two levels:

- At the *microscopic* level, we are interested in the behaviors of each individual in the spreading process, e.g. to whom and how fast does a user forward information? (Sec. 4)
- At the *macroscopic* level, we ask what are the structural properties of the spreading processes? And what is the best model for the observed structures? (Sec. 5)

At the microscopic level, we find that information spreading is indeed highly dependent on social context as well as the individuals' behavioral profiles. Macroscopically, however, we find that the tree structures observed in the spreading process can be accurately captured by a simple stochastic branching model, indicating the macroscopic structures of spreading processes, i.e., to how many people a user forwards the information and the overall coverage of the information, are largely independent of context and follow a simple reproducible pattern. To the best of our knowledge, this work presents the first comprehensive analysis of the determining factors affecting information spreading processes. We believe our findings are of fundamental importance in developing prediction models for information flow, provide new insights towards the design of our social and collaborative applications, such as assisting users to disseminate information more efficiently, protecting digital information leakage, and promoting spreading strategies to achieve expected coverage.

2. RELATED WORK

In this section, we review three categories of related work: studies on information spreading and cascades, social network analysis especially on emails, and virus propagation.

Information Spreading and Cascades. Various studies in online domains have been conducted to understand the structural properties of information flow. Among them, the spreading processes of specific pieces of information, including studies on internet chain letters and viral marketing, are most related to our work. Liben-Nowell and Kleinberg [27] studied information flows on a global scale using internet chain letters. They found that the structures of observed trees are narrow and deep. They proposed a probabilistic model, leveraging the structure of other social networks, to explain the deep tree-like structure. Golub and Jackson [13] then showed that the structures observed in [27] could be explained by the Galton-Watson branching model [35] combined with the selection bias of observing only the largest trees. Leskovec, Adamic and Huberman [25] studied an incentivized word-of-mouth effect by analyzing viral marketing data, focusing on the overall properties of the resulting recommendation network and its dynamics. By using data from a viral experiment of recommending newsletters, Iribarren and Moro [19] modeled the overall dynamics of information flow from individual activity patterns. There has been extensive work done in the blog domains about cascading behaviors [24, 2, 16, 26], and several models have been proposed to capture the structure of the blogosphere.

Previous work focuses on analyzing the observed properties of information flows. In contrast, the questions we are interested in this study are *Why does information spread? What are the factors that could potentially affect this process?*

Emails. Much work has focused on email communication records, from their static topological structure [11, 1] to dynamic properties [12, 4, 33, 22, 19]. These works focus on the overall structure of the social network, or on the timing of events. Recently, Karagiannis and Vojnovic [20] studied the behavioral patterns of email usage in a large-scale enterprise by looking at email replies. They examined various factors that could potentially affect email replies, focusing on pair-wise interactions and aiming to inform the design of advanced features. Our approach presents a new angle to using email data. First, we treat social networks as the backbone of the spreading processes, using the network topology to inspect the structures of information spreading and to assess models. Second, the spreading processes we study go beyond the pair-wise interactions of email replies, representing richer structural properties.

Virus Propagation. There is much literature regarding virus propagation. To name a few, Hethcote [18] studied the epidemic threshold for cliques. Briesemeister et al [7] studied the virus propagation on power law graphs by simulations. Most recent research has been devoted to real, arbitrary graphs. For example, Wang et al. [34] gave the analytic epidemic threshold for an arbitrary graph. Based on that, Tong et al [31] proposed an effective immunization strategy by approximately maximizing that threshold.

Virus propagation, although bearing some high level similarity to spreading of information, is not selective, meaning that a better connected individual in the network will infect more people. Information, on the other hand, spreads purposefully, representing a more complex behavior. Indeed, in our study we investigate the factors that affect to whom, and to how many people, a user forwards information, exploring the selective process of information spreading.

3. PRELIMINARIES

In this section, we describe the datasets used in our study and present the basic properties of the dataset.

3.1 Data Description

We collected detailed electronic communication records of 8,952 volunteer employees in more than 70 countries over a two-year period within a large global technology company with over 400,000 employees. Each log entry specifies the sender and receiver(s) of the message, a timestamp, the subject, and the content of the body of the email. To preserve privacy[1], the email addresses of users are hashed, and the original textual content in email's body was not saved. Instead, this content is represented as a term-frequency vector containing the terms that appear in the text as well as their counts after stemming and removal of stop-words. During the two year period, we observe about 20 million emails sent by our users. For the same population, we gathered information specifying a range of personal attributes (gender, job role, departmental affiliation and report-to relation with managers, and more). We also collected detailed financial performance data for more than 10,000 consultants in the company. These consultants generate revenue by logging

[1] refer to [37] for more information about privacy related solutions

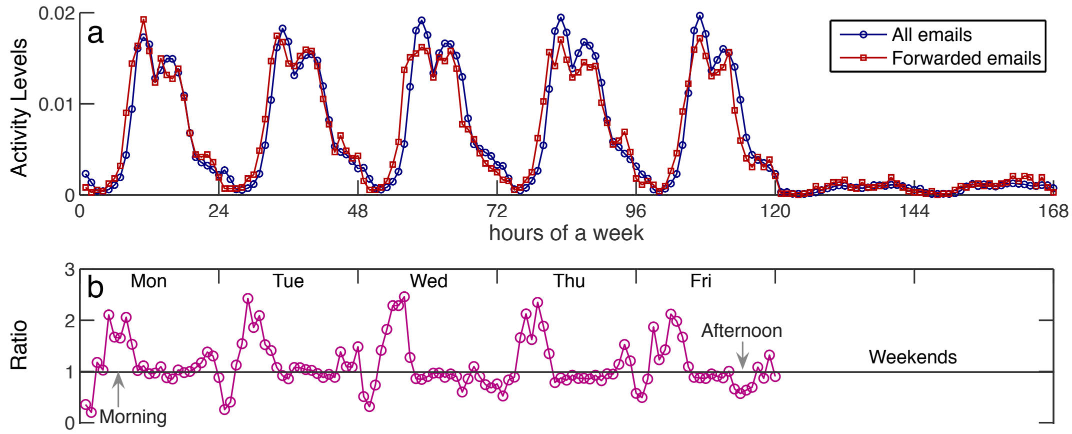

Figure 1: Activity levels at different times of the week for email forwarding activities and overall email traffic in (a), and ratios of the two in (b). The ratios in weekends are omitted due to the low volumes.

"billable hours". It has been found that a consultant's ability to generate revenue is an appropriate productivity measure [39]. Therefore, we measure performance of individuals with the total US dollars a consultant generated from June 2007 to July 2008. Combining the financial and communication data yields a total of 1,029 consultants for whom we have both email and financial records. The identities of participants were hashed.

As we are interested in how a specific piece of information spreads, we further processed our dataset using the following procedures. We started by looking for the string *Fw:* in the beginning of each email subject title. This process gives us all the emails that were forwarded. We then grouped emails with the same subject title together, reconstructing the original threads. Each forwarded thread results in an information spreading tree structure, where the single piece of information, the original body of the email, was passed from one to others. Our dataset provides us with 9,623 such distinct threads, the starting point of our study.

3.2 Basic Properties

As our dataset captures communication within an enterprise, the temporal patterns and the organizational roles of individuals involved may indicate the importance of forwarded emails. We show in Fig. 1a the activity levels, the number of communications in each hour of the week normalized by the total number of communications in a week. We observe a clearly periodic pattern. Communication builds up in the morning and decays in the afternoon with a notable dip at noon indicating the lunch time. There are two interesting points we want to make here. First, while the activity levels of forwarded emails (red squares) follow a similar periodic pattern to the overall email traffic (blue circles), their activity levels are significantly higher than the normal email traffic on workday mornings, especially on Mondays, and lower in the afternoons especially on Fridays. This can be seen clearly in Fig. 1b, where we show the ratio of the two curves in Fig. 1a. The curve goes above 1 in the mornings, but mostly below 1 in the afternoons. This is a good indicator that forwarded emails are timely and important, representing a special class of overall email traffic. Second, access to email is limited by weekly schedules. This weekly cycle becomes important when we inspect the efficiency of information spreading in the following sections. That is, there is a time delay when forwarding an email after receiving it. For example, a delay of two days in the delivery of information, when it was received on Friday, could be due to the inability of a user to access his or her email during the weekend. Therefore, for any calculation regarding time in the following sections, we perform a check by removing the off-hours. Yet no results changed qualitatively.

In addition, we observe that 38% of the forwarded threads involve people from multiple departments. This suggests that email forwarding is an important means to facilitate cross-organization collaboration. Moreover, 43% of the emails are forwarded by managers, indicating that email forwarding is a common management tool.

4. MICROSCOPIC INFORMATION SPREADING IN CONTEXT

What factors could potentially affect the information spreading process at the microscopic level, i.e., to whom and how fast a user spreads the information? Why does some information get rapidly processed and passed on to others, while other information experiences notable delay? Or more generally, why does some information get forwarded at all in the first place?

Here we investigate several aspects of these questions. Our analysis will focus on the most fundamental building blocks of information spreading – information pathways, as illustrated in Fig. 2. More specifically, user A sent an email to user B at a certain time with a specific title. Then user B waited for some time and forwarded the message to user C, passing the information, the main body of the email, along via B from A to C. We refer to user A as "initiator", B as "spreader", and C as "receiver". The dissemination process can be far more complex than this simple case, as we shall see in Section 5, where we focus on spreading processes at the macroscopic level, exploring to what extent the overall structure of spreading processes depends on context. Yet any spreading tree structures can be reduced to a combination of such information pathways.

Our study shows that not only does the social and organizational context affect to whom a specific piece of information is forwarded, but it also affects how fast it is forwarded. We found that information undergoes interesting re-routing processes, from weak links to strong ties, and from non-

Figure 2: An illustrative example of an information pathway.

experts to experts. The efficiency of the spreading process is affected by departmental structure, but little by individual performance. These findings can guide us to build better social and collaborative environments, design applications assisting users to disseminate information more efficiently, and develop strategies to protect digital information leakage and predictive tools for recommendation systems.

4.1 The Underlying Social Networks

How information spreads may be influenced by the underlying social network, and understanding the interplay between the social network and spreading process is very important. First, it has a number of implications in various social systems, such as promoting new strategies in viral marketing by taking into account the effect of the network topology. Second, it plays an important role in assessing the choice of models, arguing whether a flu-like epidemiology model, which directly relies on the topology of the network, is suitable for modeling the information flow, (see Section 5).

We start by building a social network among our users by aggregating email communications over a one year period. We add a link between two users if there has been at least one email communication between them. The weight of the link, $w(i \to j)$, is asymmetric, defined as the number of emails sent from user i to j. As we are mostly interested in the connectivity between individuals, we focus on the static picture of the network rather than the dynamics of the network evolution.

We show in Fig. 3 the probability ratio of email forwarding activity[2] as a function of the weight of the links between initiators and spreaders, $w(A \to B)$, and spreaders and receivers, $w(B \to C)$, defined in the information pathways in Fig. 2. A positive slope would indicate that information is more likely to flow through strong ties, whereas a negative slope shows that weaker connections are more favorable for the spreading processes. Surprisingly, we observe that the information is more likely to spread initially via weak ties and then gets passed through strong connections, strong evidence of information routing by spreaders choosing social neighbors of different closeness.

This raises an interesting question: how well are the information spreaders connected in the network? Are they a random sample of individuals or are they a biased sample of more central social hubs? We show in Fig. 4 the degree distribution $P(k)$ of nodes in the whole social network as grey circles and the sample of spreaders as orange crosses. Interestingly, we find that spreaders show nearly the same

[2]the probability ratio of email forwarding as a function of quantity q is obtained by $P^{\mathrm{Fw}}(q)/P^{\mathrm{rand}}(q)$, where $P^{\mathrm{Fw}}(q)$ is the probability of having q in forwarded emails, while $P^{\mathrm{rand}}(q)$ is the same probability for overall emails. A value equals to 1 would indicate $P^{\mathrm{Fw}}(q)$ is about what you would expect normally.

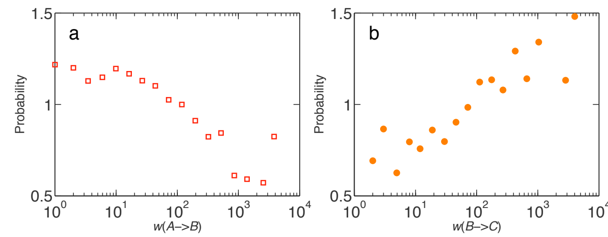

Figure 3: The probability ratio of email forwarding as a function of the weight of links between A and B and B and C respectively in the information pathways. Information spreading undergoes an interesting re-routing process, from weak links to strong ties.

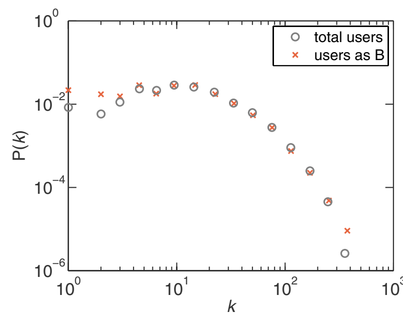

Figure 4: The degree distribution of the whole network and the group of the spreaders. The spreaders have comparable connectivity to randomly sampled individuals in the network.

distribution of connectivity as a random sample of individuals from the network.

4.2 Information Content and Expertise

An important question about information spreading is how the process depends on the relevance of the content of the information to the individual's expertise. Here, we explore this issue using the available message content. As mentioned previously, the content of each email in our dataset is represented as term frequencies. We build a vocabulary vector $\vec{v}_i = \langle s_1, s_2, ..., s_n \rangle$ for each user i by looking at the content of all the emails sent by i, where the length of $\vec{v}_i$ is the total number of meaningful words that have appeared in all emails for all users, thus is the same length for every user. The j-th element s_j is the score of the j-th word calculated by TF-IDF. The vector $\vec{v}_i$ will provide a measure of ranked "buzzwords" for user i, which serves as an indicator of the individual's expertise, since previous studies have shown emails are a primary form of communication within big corporations [36, 6]. Next, we build a vector $\vec{v}_l$ for each email l following the same procedure, where s_j is the TF-IDF score of the j-th word in $\vec{v}_l$. Therefore, $\vec{v}_l$ will give us a measure of the content of email l, accounting for overly common words and overly rare words. Then the similarity between the content of the information l and the individual i's expertise is defined as the cosine similarity of the two vectors, $\mathcal{S}_{i,l} = \vec{v}_i \cdot \vec{v}_l / (\|\vec{v}_i\| \, \|\vec{v}_l\|)$. We show in Fig. 5, how the probability ratio of information spreading changes in function of $\mathcal{S}_{i,l}$ for user i as (a) spreaders and (b) receivers, respectively. The probability ratio anti-correlates with $\mathcal{S}_{i,l}$, similarity between information content and spreaders' expertise, yet exhibits a significant positive correlation for the receivers' case. This finding offers quantitative evidence that the information undergoes a clear re-routing, demonstrating that information flows from non-experts to experts. That is, the information is more likely to be passed on by spreaders if the content is dissimilar to spreaders' expertise. It then flows to receivers who are more likely to be interested in the information.

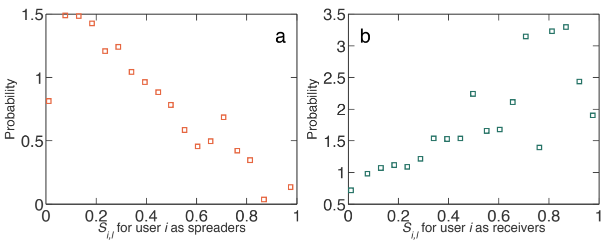

Figure 5: Probability ratio of information spreading changes in function of $\mathcal{S}_{i,l}$ for (a) spreaders and (b) receivers. Information spreading undergoes an interesting re-routing from non-experts to experts.

4.3 Organizational Context

In an enterprise, understanding how information flows within and between different departments and organizational levels is of great importance, from building a better collaborative environment to controlling information security. Here we examine the impact of organizational context in two directions: one is the influence of departmental restrictions, and the other is the organizational hierarchies.

In Fig. 6, we show the median time delay in information spreading for spreaders as different roles of brokerage [14]. There are in total five types of brokers. Figure 6 contains illustrative examples for all five: Each box represents a department, and users are from the same department if they are in the same box. If there is only one user in the department, we omit the box for brevity. Our dataset has individuals from as many as 19 departments, and the information pathways consisting of people from different departments are classified into these 5 categories. We observe that the information flows significantly faster in two cases – coordinator and gatekeeper – than the other three cases. These are the only two cases where spreaders and receivers are in the same department. Thus the bottleneck of information flow in the departmental context is to get the information out of the department. We further break down the manager and non-manager cases for each role of brokerage. We find that managers are better as a representative while non-managers are better as a liaison, but the difference between managers and non-managers is seldom large.

We now turn our attention to the impact of organizational levels. While it is intuitive to assume that users would respond faster to emails from people of higher level in the organization (e.g., the reaction of the emails are influenced by the report-to relationship), a previous study [20] on email replies revealed that the reply time does not depend on level difference. Our study shows similar results, confirming that the time delay of information appears to be independent of the hierarchy. Yet, when we look at the probability ratio of email forwarding as a function of the level difference

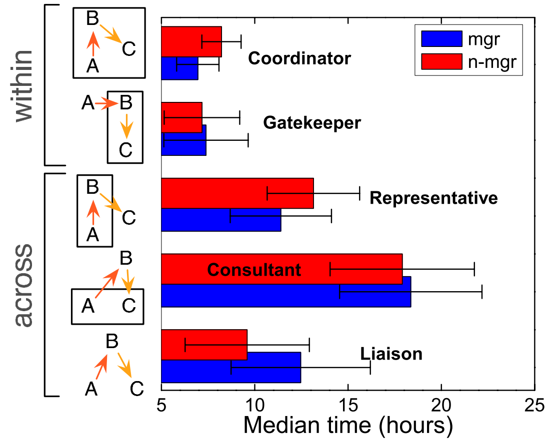

Figure 6: Information flow in the departmental context. Each box represents a department, and users are from the same department if they are in the same box. Information spreads faster when B and C are in the same department.

(Fig. 7a) and organizational distance (Fig. 7b) between initiators and spreaders, we discover some non-homophily effect as opposed to the homophily effect found in [20]. As shown in Fig. 7a, information is unlikely to flow between individuals in the same level compared with normal email traffic, and two extreme cases clearly stand out – either bottom up or top down the hierarchy. It tells us that, while the communication between different hierarchies does not yield a faster or slower response, it does matter when determining whether one would decide to pass on the information or not in the first place. Moreover, Fig. 7b further confirms the non-homophily effect that the information tends to flow between individuals at a larger distance in the formal organizational structure. This effect shows that "informal networks" and "formal networks" complement each other in information spreading.

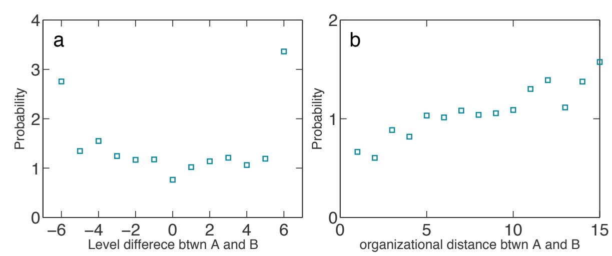

Figure 7: Probability ratio of email forwarding as a function of (a) hierarchical level difference and (b) organizational distance between initiators and spreaders. The information spreading exhibits some non-homophily effect.

4.4 Individual Characteristics

Another factor that may impact the efficiency of information spreading relates to the individual characteristics of those participating in the spreading process. Do people with different work performance behave differently in getting the word out? A natural hypothesis is that people with better performance are more efficient in spreading the information. While it is generally difficult to get a quantitative measurement of individual performance, the mentioned "billable hours" data serves this purpose. As a consultant's performance is directly related to the total revenue s/he generates, this unique data offers us an opportunity to explore for the first time how individual characteristic affects the information spreading process. To test this hypothesis, we look at whether there is a correlation between the delay time of information spreading and the performance of the individuals. We find that the hypothesis is not supported by our data. We show in Fig. 8 the correlation between the median information's waiting time in hours and the performance of initiators and spreaders, respectively. The dashed grey lines show the 25% quantiles. The information's waiting time appears to be constant for both initiators and spreaders, independent of individual performance.

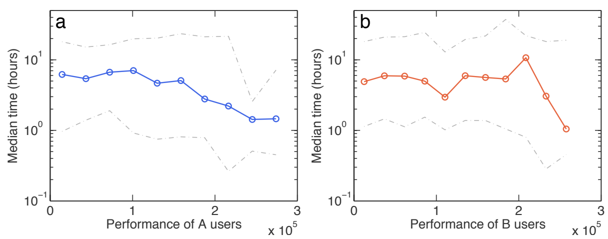

Figure 8: Information delay time in hours versus individual performance for (a) initiators and (b) spreaders. The quartiles are shown as grey lines. The efficiency of the spreading is little affected by individual performance.

5. MACROSCOPIC INFORMATION SPREADING IN CONTEXT

We now go beyond information pathways and turn our attention at the macroscopic level, aiming to understand to what degree spreading processes rely on contextual factors. That is, to how many people a user forwards the information and the information reaches in total. Our dataset contains more than 2000 threads[3]. Each thread can be treated as a rooted tree (Fig. 9),where information spreads from one user to others.

We focus on two questions: (i) what are the generic properties of the tree structures within spreading processes? (ii) how much contextual information do we need to incorporate in the models in order to capture these properties? We found, in contrast to the *narrow and deep* trees observed in previous studies [27, 13], that the trees in our study are *bushy yet shallow*. The information fans out, but quickly dies out. We further demonstrated that the way information fans out, i.e., to how many people a user forwards the information, features a high degree of randomness, being independent of the connectivity of spreaders in the underlying social network. The overall structural properties of spreading processes can be captured surprisingly well by a simple stochastic model, indicating that information spreading is largely independent of context at the macroscopic level.

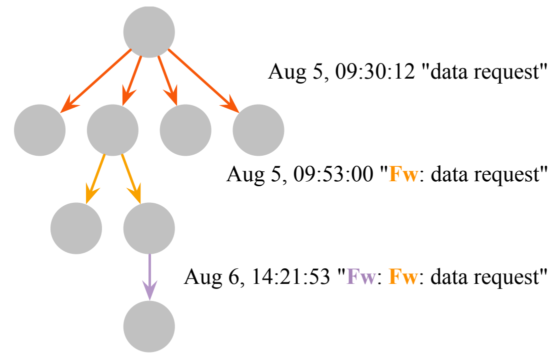

Figure 9: An illustrative example of an information spreading tree. This tree is of size 8, width 4, and depth 3.

5.1 Empirical Observations

In this subsection we report the main observations about structural properties of the threads. These observations build the foundation of our models. In summary, there are two interesting findings regarding the observed trees, which can not be interpreted intuitively by existing models.

- **Ultra-shallow trees**: Almost 95% of trees are of depth 2, and trees with more than 4 hops are absent.
- **Stage dependency**: The branching factor (number of children each node has) depends on the distance from the root.

5.1.1 Tree size, width, and depth

The size, width, and depth of a tree are its three most important structural characteristics. The size of a thread is defined as the total number of people involved in the spreading process; the width of a tree is the maximum of number of nodes in each level among all levels of the tree; depth is the length of the longest path from a leaf to the root. (As our forwarding process is conditioned on the emails that were forwarded, the minimum depth of the trees is 2). The distributions of size and width both follow a power law[4], with an exponent of 2.67 and 2.53, respectively (Fig. 10). While the power law distribution itself is not unexpected, what is surprising is that the tails of these two distributions have similar exponents. This fact directly implies, as shown in Figure 11, that the size of the trees grows almost linearly with the width (a power law relation with exponent around 1). Moreover, we observe that the tree structures extracted from email forwarding activities are ultra shallow: 95% of the trees are of depth 2, and the distribution of depth decays so fast that we don't observe any tree of depth greater than 4 within 2000+ samples.

[3]Since we did not have all the communications within the enterprise, we were left with a relatively small number of threads. The readers might be curious whether this sampling issue would affect our observations of the tree structures. As our upcoming stochastic model, which well captures the empirical observations, is purely based on the intrinsic media properties of email systems (i.e., number of recipients n in each email, and its distribution $P_n(n)$), we can therefore validate our results by checking the distribution of n across different datasets. To this end, we measured this quantity in other email datasets([12] and [11]). We found that all email datasets to date share the common feature that $P_n(n)$ universally follows a fat-tailed distribution, indicating that our results are robust to sampling.

[4]The likelihood of power law distributions and the exponents hereafter are assessed by applying the techniques in [9].

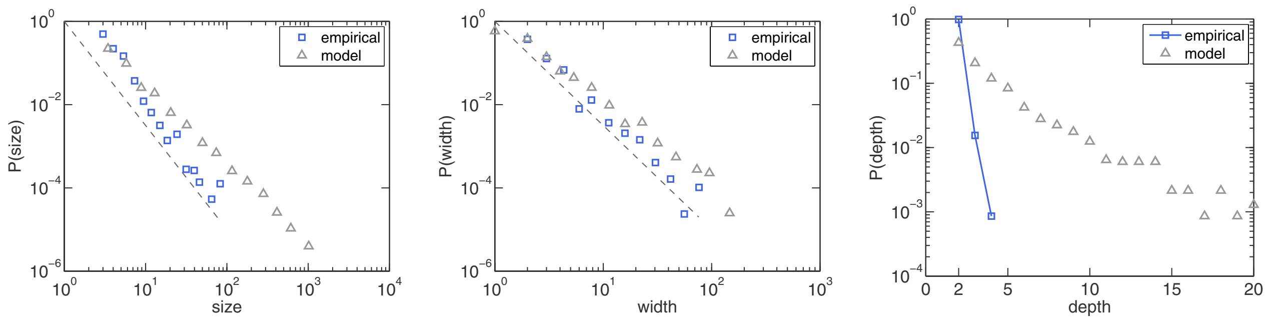

Figure 10: Distributions of size, width, and depth of the trees. Empirical measurements are denoted as blue squares, while the grey triangles are predictions of existing models. Dashed lines are guides for the eye, with an exponent of 2.5. The existing models overestimate the tails of the distributions.

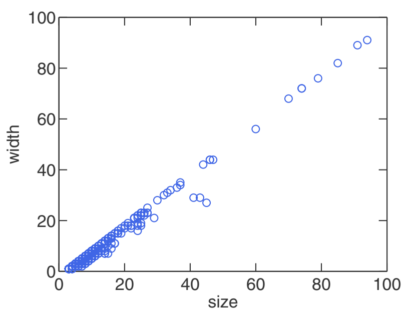

Figure 11: Scatter plot of the size and the width of the trees. The size of the tree grows almost linearly wrt the width of the tree.

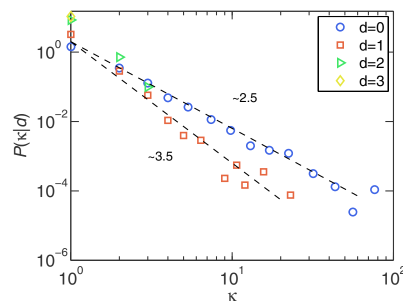

Figure 12: Distribution of branching factors κ conditioned on the distance to the root of the tree d. The branching process depends on the spreading stages.

These findings are puzzling when we apply the classical model for generating a random tree structure: a Galton-Watson branching process [35], in which each node has a random number of children κ, drawn independently according to the same distribution, denoted as $P(\kappa)$. Previous work [13] has shown that, despite the complexities of the process, this simple model fits the data quite well. We therefore follow the modeling procedure of [13] to fit our observation. We first compute the parameters $P(\kappa)$ of a Galton-Walton process by using maximum-likelihood estimation. Then we simulate the process and generate the same number of trees as empirical measurements. The distribution of size, width and depth, plotted as grey triangles in Fig. 11, follow a power law, with an exponent of 1.96 and 2.11. Clearly, directly applying this procedure significantly overestimates the tails of distributions, generating trees that are much bigger and deeper than observed empirically. Most prominent is the depth distribution. For trees that are in the subcritical regime, i.e., the mean μ of $P(\kappa)$ is less than 1 ($\mu < 1$), the depth distribution has an exponential tail [17]. However, the measured depth distribution decays much faster than the model prediction.

In summary, the trees we observed here are bushy yet very shallow, which implies that the information spreads efficiently, reaching out to many people then quickly dying out.

5.1.2 Stage dependence

The observations above raise an important question: does the information spreading process change in different stages? We therefore compute the conditional probability of κ given the distance to the root d, $P(\kappa \mid d)$, in Figure 12. In a Galton-Watson branching process, $P(\kappa)$ is universal across all nodes, therefore independent of d, predicting the collapse of curves in the plot. We observe that, however, the branching process does depend on the distance to the root. The power law exponent γ_0 of $P(\kappa)$ when $d = 0$ approximately satisfies $\gamma_1 = \gamma_0 + 1$, where γ_1 is the exponent of $P(\kappa)$ when $d = 1$. ($\gamma_0 = 2.48$ and $\gamma_1 = 3.48$). The distribution of κ becomes steeper as we move deeper down the trees, corresponding to the stage dependence, which was also observed in a recent study [23] regarding how online conversation forms yet remained unclear why the exponent of $P(\kappa)$ changes with d, indicating that this effect is generic among different settings, and a model that could appropriately capture this feature would be of great importance in enhancing our understanding of social systems.

5.2 Modeling the information spreading process

What is the underlying mechanism that governs the information spreading process? Our goal here is to explore how much contextual information we need to rely on to model the observed macroscopic structural properties of the trees in Sec. 5.1, aiming to quantify to what extent spreading processes at the macroscopic level depend on context.

The observed fat-tailed distributions of branching factors κ in Fig. 12 help us assess the properties of nodes in the

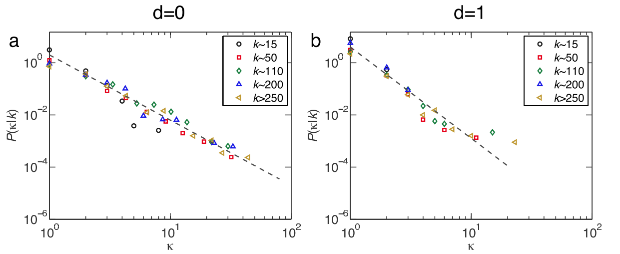

Figure 13: Distribution of branching factors κ conditioned on the degree of the node k. The branching factors are independent on the degree connectivity.

underlying social network. As shown in Fig. 4, the degree distribution, $P(k)$, is also fat tailed [5, 3, 8]. Indeed, individuals are connected differently in the network. While most people have only a few connections, there are a notable number of individuals who have many social neighbors. This raises an important question: to what extent does the information spreading process depend on the underlying social network? First off, the branching factor κ for an individual in the spreading process is upper-bounded by the total number of connections s/he has. Yet beyond that, it is important to inspect whether there is a correlation between k and κ. This question has a number of important implications. In the viral marketing case for example, where the underlying social network is usually not visible, the correlation between k and κ will tell us whether it is a good marketing strategy to carefully choose the seed populations to spread an advertisement. A positive correlation suggests that it does matter who you choose to start the spreading, as social hubs would tend to send the information to more people. Yet if the correlation is not so strong, one could argue that perhaps it is not so important how one chooses the seed population. Another example comes from the difference between the spreading of information and diseases. Indeed, diseases spread from a seed to many others through networks, bearing high level similarity to the spreading of information. The models of epidemics commonly rely on infection rates, where better connected nodes infect more neighbors, corresponding to a strong correlation between k and κ. Therefore, understanding to what extent information spreading relies on the context of underlying network would quantitatively assess the difference between these two spreading processes, arguing whether the existing epidemic models are applicable to the spreading of information.

The correlation between k and κ can be examined by empirically measuring the conditional probability $P(\kappa \mid k)$. Indeed, as $P(\kappa) = \int P(\kappa \mid k)P(k)dk$, if κ is largely uncorrelated with k, $P(\kappa \mid k)$ can therefore be factored out of the integral, giving $P(\kappa) = P(\kappa \mid k)$, leading to a data collapse when plotting $P(\kappa)$ in different curves by grouping individuals of similar k. We show in Fig. 13 the conditional probability $P(\kappa \mid k)$ for two different stages of spreading ($d = 0$ vs. $d = 1$), respectively. Surprisingly, we observe very good collapse for different k in both figures, which indicates that there is no direct correlation between k and κ. The breadth of the dissemination of information is independent of the connectivity k of individuals. This indicates, while to whom a user forwards the information indeed depends on the underlying social network (as shown in Sec. 4.1), to how many people (κ) one would forward the information does not.

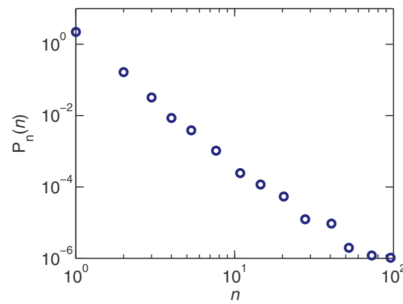

Figure 14: The distribution of the number of recipients for each email $P_n(n)$ is fat-tailed.

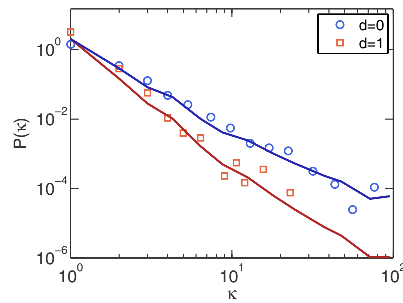

Figure 15: $P(\kappa)$ for $d = 0$ and $d = 1$, model prediction (solid lines) vs. experiment measure (scattered squares and circles). Our model well captures the stage dependence phenomenon of information spreading.

The surprising independence of node properties of the information spreading process leads us to question its dependence on the media properties of email systems. Therefore, we model the spreading processes by mimicking the way emails are sent. Indeed, an important feature of email communication, distinguishing it from other forms of communication, like cell phones, is the ability to send a message to multiple recipients at the same time.

Therefore, the distribution of the number of recipients for all the emails being sent should follow some non-trivial form, other than $\delta(1)$ in cell phones, i.e., each phone call is made to one recipient only. Let us denote the distribution for emails system as $P_n(n)$ for now, where n represents the number of recipients in each email. While some emails are forwarded, many more are not. The easiest way to look at email forwarding is to treat it as an independent decision making process, where each recipient with probability p forwards the information, or probability $1-p$ does nothing. As email forwarding represents a small fraction of overall email traffic, p should be a small number. When a recipient decides to forward the email, s/he draws a random number from the distribution P_n to decide how many people to send the emails. So the distribution of branching factors should follow the same distribution as P_n, from which the random numbers were drawn, giving $P(\kappa) = P_n(\kappa)$. However, this should only hold for the case of $d > 0$. Indeed, as our study is focused on the emails that are forwarded, there should be an extra term for correcting this conditional probability when $d = 0$. That is, the original emails with more recipi-

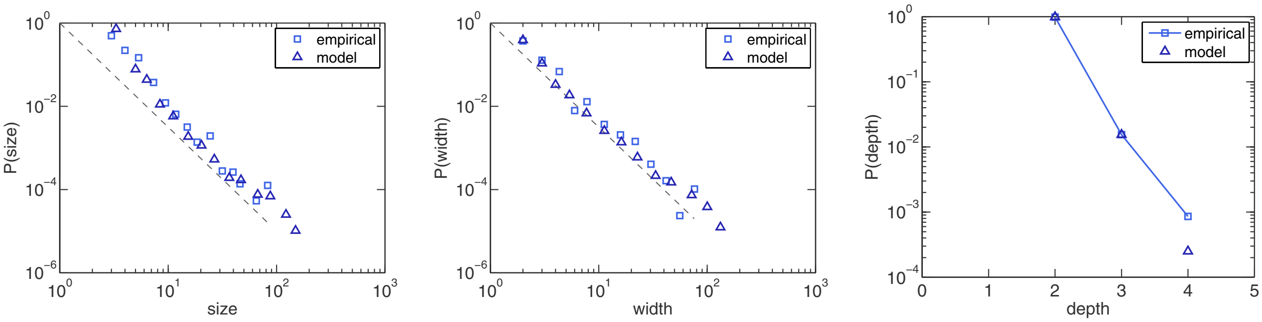

Figure 16: Size, width, and depth distributions of model prediction (triangles) with empirical observations (squares). The model matches well with observations. Note the last point in depth distribution is biased by empirical finite size effect, lower bounded by N^{-1}.

ents are more likely to get forwarded, as there will be more people to make a decision whether or not to pass on the information. Following this mechanism, the distribution of branching factors at depth 0, $P(\kappa \mid d = 0)$, follows

$$\begin{aligned} P(\kappa \mid d = 0) &= A\left(1 - (1-p)^{\kappa}\right) P_n(\kappa) \\ &= A\left(1 - e^{\kappa \ln(1-p)}\right) P_n(\kappa) \end{aligned} \tag{1}$$

where A is the normalization factor, whereas $P(\kappa \mid d > 0)$ follows

$$P(\kappa \mid d > 0) = P_n(\kappa) \tag{2}$$

In the limit of $p \to 0$, to the leading power, the relationship between the scaling exponent γ_0 of $P(\kappa \mid d = 0)$, and γ_1 of $P(\kappa \mid d = 1)$, follows the simple relation $\gamma_1 = \gamma_0 + 1$ if $\kappa \ll -1/\ln(1-p) \approx 1/p$, and $\gamma_1 = \gamma_0$ if $\kappa \gg -1/\ln(1-p) \approx 1/p$.

Both parameter p and function P_n can be measured independently from our data, yielding $p = 0.012$ and a fat-tailed distribution P_n (Fig. 14). We can therefore simulate the distributions of size, width, and depth using these two measured parameters. The results are shown in Fig. 16, with observations as squares and model predictions as triangles. Surprisingly, they all match the empirical observations very well. The distribution of size and width follows a power law, with an exponent of 2.63 and 2.51, very close to the empirical observations (2.67 and 2.53). Indeed, both distributions pass the two-sample Kolmogorov-Smirnov tests, with p-values equal to 1. Furthermore, the observation of stage dependence could be verified analytically by plugging the parameters into eqs. (1) and (2), as plotted in blue and red lines in Fig. 15, respectively. It is also very well captured by the model.

The model we described above for email forwarding processes is purely stochastic and has two parameters, p and P_n, which are measured from our email dataset independently. Perhaps unexpectedly, such a simple model explains a great deal of observations. This, together with Fig. 13, indicates that, despite the complexity in real life, the macroscopic structures of information spreading processes are largely independent of contextual information and can be well captured and explained via simple machanisms.

6. CONCLUSIONS AND FUTURE WORK

Applications of social systems rely on our understanding of information spreading patterns. In this work, by combining two related but distinct large scale datasets, we address the factors that govern information spreading at both microscopic and macroscopic levels. We found, microscopically, whom the information flows to indeed depends on the structure of the underlying social network, individual expertise and organizational hierarchy. The performance of individuals has little influence on the efficiency of spreading, yet departmental constraints do slow down the process. At the macroscopic level, however, although seemingly complex, the structural properties of spreading trees, i.e., to how many people a user forwards the information and the total coverage the information reaches, can be well captured by a simple stochastic branching model, indicating that the spreading process follows a random yet reproducible pattern, largely independent of context. We believe that our findings could guide users to build better social and collaborative applications, design tools and strategies to spread information more efficiently, improve information security, develop predictive tools for recommendation systems, and more.

Future directions mainly fall into two lines. The first is to develop a better prediction model for information flow. Indeed, upon understanding to whom one forwards information, when one would forward it, and to how many people, the question thereafter is can we build a better prediction model of the flows? The second direction is about the mutation of information. People sometimes add extra information or express opinions about existing information when passing along the originals to others. How does information mutate along the way? How does the mutation of information affect the patterns of spreading? These questions stand as missing chapters in our understanding of spreading processes. Indeed, with the availability of large-scale email datasets, thorough inspection of the email message contents will reveal the dynamics of information itself, which in turn can yield better predictive tools for information spreading.

Acknowledgement The authors thank James P. Bagrow, Nick Blumm, Helena Buhr, U Kang, Yu-Ru Lin, Yelena Mejova, Jiang Yang, and anonymous reviewers for many useful discussions and insightful suggestions. This work was supported by the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Agreement Number W911NF-09-2-0053; the James S. McDonnell Foundation 21st Century Initiative in Studying Complex Systems; the US Office of Naval Research Award (N000141010968); the NSF within the Information

Technology Research (DMR-0426737), and IIS-0513650 programs and the Defense Threat Reduction Agency Award HDTRA1-08-1-0027.

7. REFERENCES

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187 – 203, 2005.

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

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97, 2002.

[4] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[5] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[6] N. K. Baym, Y. B. Zhang, and M.-C. Lin. Social Interactions Across Media. *New Media & Society*, 6(3):299–318, 2004.

[7] L. Briesemeister, P. Lincoln, and P. Porras. Epidemic profiles and defense of scale-free networks. *WORM 2003*, Oct. 27 2003.

[8] G. Caldarelli. *Scale-Free Networks*. Oxford University Press, 2007.

[9] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[10] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.

[11] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66(3):035103, Sep 2002.

[12] J.-P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, 2004.

[13] B. Golub and M. Jackson. Using selection bias to explain the observed structure of Internet diffusions. *Proceedings of the National Academy of Sciences*, 107(24):10833, 2010.

[14] R. Gould and R. Fernandez. Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, 19(1989):89–126, 1989.

[15] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, May, 1973.

[16] D. Gruhl, R. V. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501, 2004.

[17] T. Harris. *The theory of branching processes*. Dover Pubns, 2002.

[18] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599–653, 2000.

[19] J. L. Iribarren and E. Moro. Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.*, 103(3):038702, Jul 2009.

[20] T. Karagiannis and M. Vojnovic. Behavioral profiles for advanced email features. In *WWW*, pages 711–720, 2009.

[21] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[22] G. Kossinets and D. J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90, 2006.

[23] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *KDD*, pages 553–562, 2010.

[24] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW*, pages 159–178, 2005.

[25] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *ACM/EC*, pages 228–237, 2006.

[26] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, 2007.

[27] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[28] H. Smith, Y. Rogers, and M. Brady. Managing one's social network: Does age make a difference. In *In: Proc. Interact 2003, IOS*, pages 551–558. Press, 2003.

[29] M. A. Smith, J. Ubois, and B. M. Gross. Forward thinking. In *CEAS*, 2005.

[30] D. Strang and S. A. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24(1):265–290, 1998.

[31] H. Tong, B. A. Prakash, C. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, and D. H. P. Chau. On the vulnerability of large graphs. In *ICDM*, 2010.

[32] T. W. Valente. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory*, 2:163–164, 1996.

[33] A. Vázquez, J. Oliveira, Z. Dezsö, K. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):36127, 2006.

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

[35] H. Watson and F. Galton. On the probability of the extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland*, pages 138–144, 1875.

[36] B. Wellman. *The Internet in Everyday Life (The Information Age)*. Blackwell Publishers, December 2002.

[37] Z. Wen and C.-Y. Lin. On the quality of inferring interests from social neighbors. In *KDD*, pages 373–382, 2010.

[38] F. Wu, B. A. Huberman, L. A. Adamic, and J. R. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327 – 335, 2004.

[39] L. Wu, C.-Y. Lin, S. Aral, and E. Brynjolfsson. Value of social network – a large-scale analysis on network structure impact to financial revenue of information technology consultants. In *The Winter Conference on Business Intelligence*, 2009.