Information Flow in Social Groups

Fang Wu,[1] Bernardo A. Huberman,[2,*] Lada A. Adamic,[2] and Joshua R. Tyler[2]
[1] *Physics Department, Stanford University*
[2] *Information Dynamics Lab, HP Laboratories, 1501 Page Mill Road, CA 94304-1126*

We present a study of information flow that takes into account the observation that an item relevant to one person is more likely to be of interest to individuals in the same social circle than those outside of it. This is due to the fact that the similarity of node attributes in social networks decreases as a function of the graph distance. An epidemic model on a scale-free network with this property has a finite threshold, implying that the spread of information is limited. We tested our predictions by measuring the spread of messages in an organization and also by numerical experiments that take into consideration the organizational distance among individuals.

The problem of information flows in social organizations is relevant to issues of productivity, innovation and the sorting out of useful ideas out of the general chatter of a community. How information spreads determines the speed with which individuals can act and plan their future activities. In particular, email has become the predominant means of communication in the information society. It pervades business, social and scientific exchanges and as such it is a highly relevant area for research on communities and social networks. Not surprisingly, email has been established as an indicator of collaboration and knowledge exchange [1–6]. Email is also a good medium for research because it provides plentiful data on personal communication in an electronic form.

Since individuals tend to organize both formally and informally into groups based on their common activities and interests, the way information spreads is affected by the topology of the interaction network, not unlike the spread of a disease among individuals. Thus one would expect that epidemic models on graphs are relevant to the study of information flow in organizations. In particular, recent work on epidemic propagation on scale-free networks found that the threshold for an epidemic is zero, implying that a finite fraction of the graph becomes infected for arbitrarily low transmission probabilities [7–9]. The presence of additional network structure was found to further influence the spread of disease on scale-free graphs [10–12].

There are, however, differences between information flows and the spread of viruses. While viruses tend to be indiscriminate, infecting any susceptible individual, information is selective and passed by its host only to individuals the host thinks would be interested in it. The information any individual is interested in depends strongly on their characteristics. Furthermore, individuals with similar characteristics tend to associate with one another, a phenomenon known as homophily [13–15]. Conversely, individuals many steps removed in a social network on average tend not to have as much in common, as shown in a study [16] of a network of Stanford student homepages and illustrated in Figure 1.

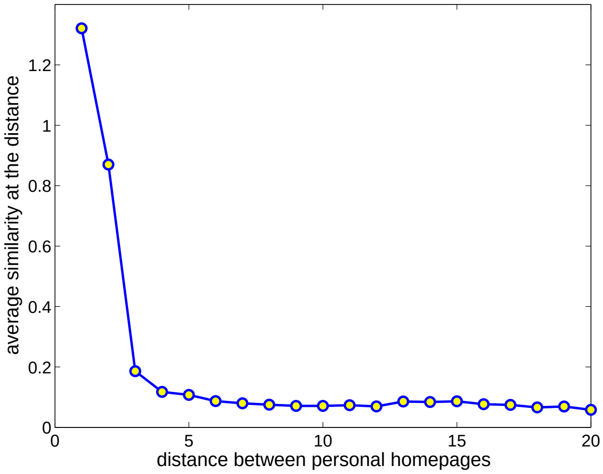

FIG. 1: Average textual similarity of Stanford student homepages as a function of the number of hyperlinks separating them. The textual similarity is measured as $\sum_x 1/\ln(f(x))$, where x are shared words and phrases such as persons, proper nouns, places, and organizations, and $f(x)$ is the number of homepages mentioning x.

We therefore introduce an epidemic model with decay in the transmission probability of a particular piece of information as a function of the distance between the originating source and the current potential target. In the following analysis, we show that this epidemic model on a scale-free network has a finite threshold, implying that the spread of information is limited. We further tested our predictions by observing the prevalence of messages in an organization and also by numerical experiments that take into consideration the organizational distance among individuals.

Consider the problem of information transmission in a power-law network whose degree distribution is given by

$$p_k = Ck^{-\alpha}e^{-k/\kappa}, \tag{1}$$

where $\alpha > 1$, there is an exponential cutoff at κ and C is determined by the normalization condition. A real world graph will at the very least have cutoff at the max-

[*] Electronic address: huberman@hpl.hp.com

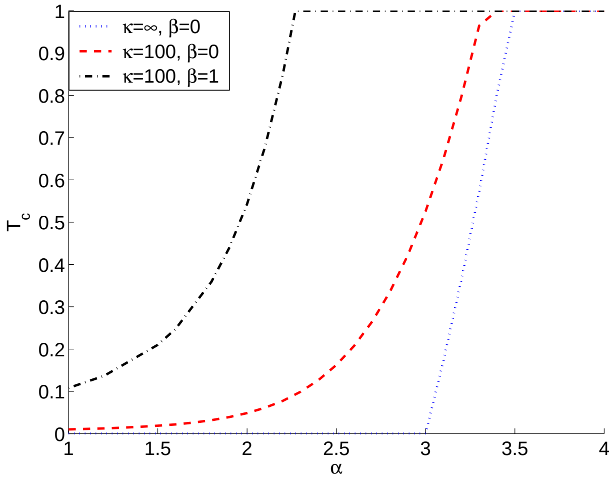

FIG. 2: T_c as a function of α. The three different curves, from bottom to top are: 1) no decay in transmission probability, no exponential cutoff in the degree distribution ($\kappa = \infty, \beta = 0$). 2) $\kappa = 100, \beta = 0$, 3) $\kappa = 100, \beta = 1$.

imum degree $k = N$, where N is the number of nodes, and many networks show a cutoff at values much smaller than N. For our analysis, we will make use of generating functions, whose application to graphs with arbitrary degree distributions is discussed in [17]. The generating function of the distribution is

$$G_0(x) = \sum_{k=1}^{\infty} p_k x^k = \frac{\mathrm{Li}_\alpha(xe^{-k/\kappa})}{\mathrm{Li}_\alpha(e^{-1/\kappa})}. \tag{2}$$

where $Li_n(x)$ is the nth polylogarithm of x.

Following the analysis in [18] for the SIR (susceptible, infected, removed) model, we now estimate the probability $p_l^{(1)}$ that the first person in the community who has received a piece of information will transmit it to l of their neighbors. Using the binomial distribution, we find

$$p_l^{(1)} = \sum_{k=l}^{\infty} p_k \binom{k}{l} T^l (1-T)^{k-l}, \tag{3}$$

where the superscript "(1)" refers to first neighbors, those who received the information directly from the initial source. The *transmissiblity* T is the average total probability that an infective individual will transmit an item to a susceptible neighbor and is derived in [18] as a function of r_{ij}, the rate of contacts between two nodes, and τ_i, the time a node remains infective. If r_{ij} and τ_i are iid randomly distributed according to the distributions $P(r)$ and $P(\tau)$, then the item will propagate as if all transmission probabilities are equal to a constant T.

$$T = \langle T_{ij} \rangle = 1 - \int_0^\infty dr d\tau P(r) P(\tau) e^{-r\tau} \tag{4}$$

The generating function for $p_l^{(1)}$ is given by

$$\begin{aligned} G^{(1)}(x) &= \sum_{l=0}^{\infty}\sum_{k=l}^{\infty} p_k \binom{k}{l} T^l (1-T)^{k-l} x^l & (5)\\ &= G_0(1+(x-1)T) = G_0(x;T). & (6)\end{aligned}$$

Suppose the transmissibility decays as a power of the distance from the initial source. We choose this weakest form of decay as the results that are obtained from it will also be valid for stronger functional forms. Then the probability that an mth neighbor will transmit the information to a person with whom he has contact is given by

$$T^{(m)} = (m+1)^{-\beta} T, \tag{7}$$

where $\beta > 0$ is the decay constant. $T^{(m)} = T$ at the originating node ($m = 0$) and decays to zero as $m \to \infty$.

The generating function for the transmission probability to 2nd neighbors can be written as

$$G^{(2)}(x) = \sum_k p_k^{(1)} [G_1^{(1)}(x)]^k = G^{(1)}(G_1^{(1)}(x)), \tag{8}$$

where

$$G_1^{(1)}(x) = G_1(x; 2^{-\beta} T) = G_1(1+(x-1)2^{-\beta} T) \tag{9}$$

and

$$G_1(x) = \frac{\sum_k k p_k x^k}{x \sum_k k p_k} = \frac{G_0'(x)}{G_0'(1)} \tag{10}$$

is the generating function of the degree distribution of a vertex reached by following a randomly chosen edge, not counting the edge itself [17]. Similarly, if we define $G^{(m)}(x)$ to be the the generating function for the number of mth neighbors affected, then we have

$$G^{(m+1)}(x) = G^{(m)}(G_1^{(m)}(x)) \quad \text{for } m \geq 1, \tag{11}$$

where

$$G_1^{(m)}(x) = G_1(x; (m+1)^{-\beta} T) = G_1(1+(x-1)(m+1)^{-\beta} T). \tag{12}$$

Or, more explicitly,

$$G^{(m+1)}(x) = G^{(1)}(G_1^{(1)}(G_1^{(2)}(\cdots G_1^{(m)}(x)))). \tag{13}$$

The average number z_{m+1} of $(m+1)$th neighbors is

$$z_{m+1} = {G^{(m+1)}}'(1) = {G_1^{(m)}}'(1) {G^{(m)}}'(1) = {G_1^{(m)}}'(1) z_m. \tag{14}$$

So the condition that the size of the outbreak (the number of affected individuals) remains finite is given by

$$\frac{z_{m+1}}{z_m} = {G_1^{(m)}}'(1) < 1, \tag{15}$$

or

$$(m+1)^{-\beta} T G_1'(1) < 1. \tag{16}$$

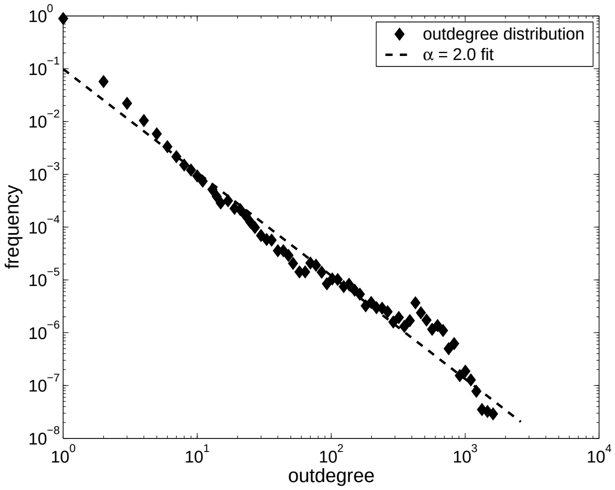

FIG. 3: Outdegree distribution for all senders (224,514 in total) sending email to or from the HP Labs email server over the course of 3 months. The outdegree of a node is the number of correspondents the node sent email to.

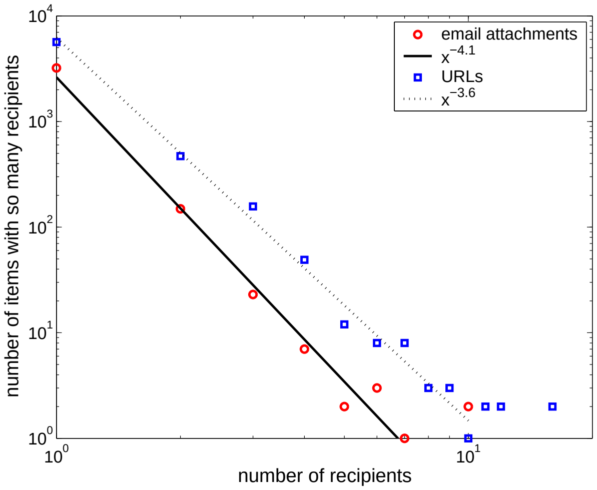

FIG. 4: Number of people receiving URLs and attachments

Note that $G_1'(1)$ does not diverge when $\alpha < 3$ due to the presence of a cutoff at κ. For any given T, the left hand side of the inequality above goes to zero when $m \to \infty$, so the condition is eventually satisfied for large m. Therefore the average total size

$$\langle s \rangle = \sum_{m=1}^{\infty} z_m \tag{17}$$

is always finite if the transmissibility decays with distance.

To compare this result with previous results on disease spread on scale-free networks, we take as an example a network made up of 10^6 vertices. We can define an epidemic to be an outbreak affecting more than 1% or 10^4 vertices. Thus for fixed α, κ and β, we can define T_c as the transmissibility above which $\langle s \rangle$ would be made to exceed 10^4.

Figure 2 shows the numerical results of the variation of T_c as a function of α. When $\beta = 0$ (there is no decay in transmission probability), $\kappa = \infty$, and $\alpha < 3$, T_c is zero and epidemics encompassing more than 10^4 vertices occur for arbitrarily small T, as was found in [8]. Keeping β at zero and adding a cutoff at $\kappa = 100$ produces a non-zero critical transmissibility T_c, as was found in [18]. For $\alpha = 2$, a typical value for real-world networks, T_c is still very near zero, meaning that for most values of T, epidemics do occur. However, when we impose a decay in transmissibility by setting β to 1, T_c rises substantially. For example, T_c jumps to 0.54 at $\alpha = 2$ and rises rapidly to 1 as α increases further, implying that the information may not spread over the network.

In order to validate empirically that the spread of information within a network of people is limited, and hence distinct from the spread of a virus, we gathered a sample from the mail clients of 40 individuals (30 within HP Labs, and 10 from other areas of HP, other research labs, and universities). Each volunteer executed a program that identified URLs and attachments in the messages in their mailboxes, as well as the time the messages were received. This data was cryptographically hashed to protect the privacy of the users. By analyzing the message content and headers, we restricted our data to include only messages which had been forwarded at least one time, thereby eliminating most postings to mailing lists and more closely approximating true inter-personal information spreading behavior. The median number of messages in a mailbox in our sample is 2200, indicating that many users keep a substantial portion of their email correspondence. Although some messages may have been lost when users deleted them, we assume that a majority of messages containing useful information had been retained.

Figure 4 shows a histogram of how many users had received each of the 3401 attachments and 6370 URLs. The distribution shows that only a small fraction (5% of attachments and 10% of URLs) reached more than 1 recipient. Very few (41 URLs and 6 attachments) reached more than 5 individuals, a number which, in a sample of 40, starts to resemble an outbreak. In follow-up discussions with our study subjects, we were able to identify the content and significance of most of these messages. 14 of the URLs were advertisements attached to the bottom of an email by free email services such as Yahoo and MSN. These are in a sense viral, because the sender is sending them involuntarily. It is this viral strategy that was responsible for the rapid buildup of the Hotmail free email service user base. 10 URLs pointed to internal HP project or personal pages, 3 URLs were for external commercial or personal sites, and the remaining 14 could not be identified.

In our sample, one group is overrepresented, allowing

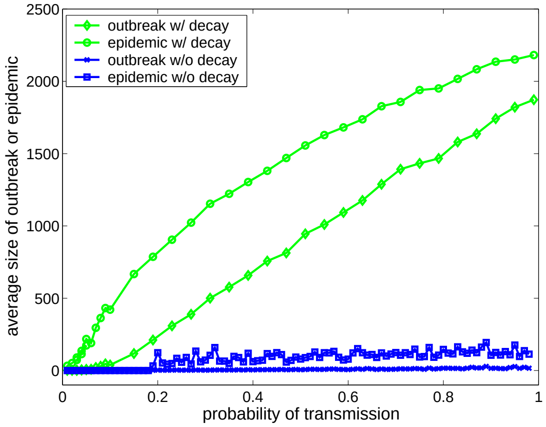

FIG. 5: Average outbreak and epidemic size as a function of the transmission probability p. The total number of potential recipients is 7119.

us to observe both the spread of information within a close group, and the lack of information spread across groups. A number of attachments reaching four or more people were resumes circulated within one group. A few attachments were announcements passed down by higher level management. This kind of top down transmission within an organization is another path through which information can be efficiently disseminated.

Next we simulated the effect of decay in the transmission probability on the email graph at HP Labs in Palo Alto, CA. The graph was constructed from recorded logs of all incoming and outgoing messages over a period of 3 months. The graph has a nearly power-law out degree distribution, shown in Figure 3, including both internal and external nodes. Because all of the outgoing and incoming contacts were recorded for internal nodes, their in and out degrees were higher than for the external nodes for which we could only record the email they sent to and received from HP Labs. We however considered a graph with the internal and external nodes mixed (as in [1]) to specifically demonstrate the effect of a decay on the spread of email in a power-law graph.

We simulated the spread of an epidemic by selecting a random initial sender to infect and following the email log containing 120,000 entries involving over 7,000 recipients in the course of a week. Every time an infective individual was recorded as sending an email to someone else, they had a constant probability p of infecting the recipient. Hence individuals who email more often have a higher probability of infecting. We also assume that an individual remains infective (willing to transmit a particular piece of information) for a period of $\tau = 24$ hours.

Next we introduced a decay in the one-time transmission probability p_{ij} as $p * d_{ij}^{-1.75}$, where d_{ij} is the distance in the organizational hierarchy between individuals i and j. This exponent roughly corresponds to the decay in similarity between homepages shown in Figure 1. Here $r_{ij} = p_{ij} * f_{ij}$, where f_{ij} is the frequency of communication between the two individuals, obtained from the email logs. The decay in transmission probability represents the fact that individuals closer together in the organizational hierarchy share more common interests. Individuals have a distance of one to their immediate superiors and subordinates and to those they share a superior with. The distance between someone within HP labs and someone outside of HP labs was set to the maximum hierarchical distance of 8.

In figure 5 we show the average outbreak size, and the average epidemic size (chosen to be any outbreak affecting more than 30 individuals) as a function of the one time transmission probability p. Without decay, the epidemic threshold falls below $p = 0.01$. With decay, the threshold is set back to $p = 0.20$ and the outbreak epidemic size is limited to about 50 individuals, even for $p = 1$.

As these results show, the decay of similarity among members of a social group has strong implications for the propagation of information among them. In particular, the number of individuals that a given email message reaches is very small, in contrast to what one would expect on the basis of a virus epidemic model on a scale free graph. The implication of this finding is that merely discovering hubs in a community network is not enough to ensure that information originating at a particular node will reach a large fraction of the community. We expect that these findings are also valid with other means of social communication, such as verbal exchanges, telephony and instant messenger systems.

[1] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103 (2002).

[2] B. Wellman, Science **293**, 2031 (2002).

[3] S. Whittaker and C. Sidner, in *Proceedings of CHI'96 Conference on Computer Human Interaction* (Logos Verlag, New York, 21996), pp. 276–283.

[4] R. Guimerà *et al.*, Physical Review E **65**, 065103 (2003).

[5] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, in *Proceedings of the International Conference on Communities and Technologies* (Kluwer Academic Publishers, Netherlands, 2003).

[6] J.-P. Eckmann, E. Moses, and D. Sergi, `http://xyz.lanl.gov/abs/cond-mat/0304433"` (unpublished).

[7] Z. Dezso and A.-L. Barabasi, Phys. Rev. E **65**, 055103 (2002).

[8] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).

[9] M. E. J. Newman, S. F., and J. Balthrop, Phys. Rev. E **66**, 035101 (2002).

[10] V. M. Eguiluz and K. Klemm, Phys. Rev. Lett. **89**, 108701 (2002).

[11] A. Vazquez *et al.*, Physical Review E **67**, 046111 (2003).

[12] M. E. J. Newman, Phys. Rev. Lett **89**, 208701 (2002).

[13] P. Lazarsfeld and R.K.Merton, in *Freedom and Control in Modern Society*, edited by M. Berger, T. Abel, and C. Page (Van Nostrand, New York, 1954), Chap. Friendship as a social Process: A Substantive and Methodological Analysis.

[14] J. Touhey, Sociometry **37**, 363 (1974).

[15] S. Feld, American Journal of Sociology **86**, 1015 (1981).

[16] L. A. Adamic and E. Adar, Social Networks **25**, 211 (2003).

[17] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).

[18] M. Newman, Phys. Rev. E **66**, 016128 (2002).