Information Dynamics in the Networked World

Bernardo A. Huberman and Lada A. Adamic

HP Labs, 1501 Page Mill Road, Palo Alto CA 94304, USA

Abstract. We review three studies of information flow in social networks that help reveal their underlying social structure, how information spreads among them and why small world experiments work.

1 Introduction

The problem of information flows in social organizations is relevant to issues of productivity, innovation and the sorting out of useful ideas from the general chatter of a community. How information spreads determines the speed by which individuals can act and plan their future activities. Moreover, information flows take place within social networks whose nature is sometimes difficult to establish. This is because the network itself is sometimes different from what one would infer from the formal structure of the group or organization.

The advent of email as the predominant means of communication in the information society now offers a unique opportunity to observe the flow of information along both formal and informal channels. Not surprisingly, email has been established as an indicator of collaboration and knowledge exchange [1–5]. Email is also a good medium for social network research because it provides plentiful data on personal communication in an electronic form. This volume of data enables the discovery of shared interests and relationships where none were previously known [6].

In this chapter we will review three studies that utilized networks exposed by email communication. In all three studies, the networks analyzed were derived from email messages sent through the Hewlett Packard Labs email server over the period of several months in 2002 and 2003. The first study, by Tyler et al. [4], develops an automated method applying a betweenness centrality algorithm to rapidly identify communities, both formal and informal, within the network. This approach also enables the identification of leadership roles within the communities. The automated analysis was complemented by a qualitative evaluation of the results in the field.

The second study, by Wu et al. [7] analyzes email patterns to model information flow in social groups, taking 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 dis-

B.A. Huberman and L.A. Adamic, Information Dynamics in the Networked World, Lect. Notes Phys. **650**, 371–398 (2004)
http://www.springerlink.com/

tance. An epidemic model on a scale-free network with this property has a finite threshold, implying that the spread of information is limited. These predictions were tested by measuring the spread of messages in an organization and also by numerical experiments that take into consideration the organizational distance among individuals.

Since social structure affects the flow of information, knowledge of the communities that exist within a network can also be used for navigating the networks when searching for individuals or resources. The study by Adamic and Adar [8], does just this, by simulating Milgram's small world experiment on the HP Labs email network. The small world experiment has been carried out a number of times over the past several decades, each time demonstrating that individuals passing messages to their friends and acquaintances can form a short chain between two people separated by geography, profession, and race. While the existence of these chains has been established, how people are able to navigate without knowing the complete social networks has remained an open question. Recently, models have been proposed to explain the phenomenon, and the work of Adamic and Adar is a first study to test the validity of these models on a social network.

2 Email as Spectroscopy

Communities of practice are the informal networks of collaboration that naturally grow and coalesce within and outside organizations. Any institution that provides opportunities for communication among its members is eventually threaded by communities of people who have similar goals and a shared understanding of their activities [9]. These communities have been the subject of much research as a way to uncover the reality of how people find information and execute their tasks. (for example, see [10–12], or for a survey see [13]).

These informal networks coexist with the formal structure of the organization and serve many purposes, such as resolving the conflicting goals of the institution to which they belong, solving problems in more efficient ways [14], and furthering the interests of their members. Despite their lack of official recognition, informal networks can provide effective ways of learning, and with the proper incentives actually enhance the productivity of the formal organization [15–17].

Recently, there has been an increased amount of work on identifying communities from online interactions (a brief overview of this work can be found in [1]). Some of this work finds that online relationships do indeed reflect actual social relationships, thus adding effectively to the "social capital" of a community. Ducheneaut and Bellotti [18] conducted in-depth field studies of email behavior, and found that membership in email communities is quite fluid and depends on organizational context. Mailing lists and personal web pages also serve as proxies for social relationships [19], and the communities identified from these online proxies resemble the actual social communities of the represented individuals. Because of the demonstrated value of communities of practice, a fast, accurate method of identifying them is desirable.

Classical practice is to gather data from interviews, surveys, or other field-work and to construct links and communities by manual inspection (see [20,21] or an Internet-centric approach in [22]). These methods are accurate but time-consuming and labor-intensive, prohibitively so in the context of a very large organization. Alani et al. [23] recently introduced a semi-automated utility that uses a simple algorithm to identify nearest neighbors to one individual within a university department.

The method of Tyler et al. [4] uses email data to construct a network of correspondences, and then discovers the communities by partitioning this network. It was applied to a set of over one million email messages collected over a period of roughly two months at HP Labs in Palo Alto, an organization of approximately 400 people. The only pieces of information used from each email are the names of the sender and receiver (i.e., the "to:" and "from:" fields), enabling the processing of a large number of emails while minimizing privacy concerns.

The method was able to identify small communities within the organization, and the leaders for those communities, in a matter of hours, running on a standard Linux desktop PC. This experiment was followed by a qualitative evaluation of the experimental results in the "field", which consisted of sixteen face-to-face interviews with individuals in HP Labs. The interviews validated the results obtained by the automated process, and provided interesting perspectives on the communities identified. We describe the results in more detail below.

2.1 Identifying Communities

It is straightforward to construct a graph based on email data, in which vertices represent people and edges are added between people who exchanged at least a threshold number of email messages. Next, one can identify communities: subsets of related vertices, with many edges connecting vertices of the same subset, but few edges lying between subsets [24].

The method of Wilkinson and Huberman [25], related to the algorithm of Girvan and Newman [24], partitions a graph into discrete communities of nodes and is based on the idea of betweenness centrality, or betweenness, first proposed by Freeman [26]. The betweenness of an edge is defined as the number of all-pair shortest paths that traverse it. This property distinguishes inter-community edges, which link many vertices in different communities and have high betweenness, from intra-community edges, whose betweenness is low.

To illustrate the community discovery process, consider the small graph shown in Fig. 1. This graph consists of two well-defined communities: the four vertices denoted by squares, including vertex A, and the nine denoted by circles, including vertex B. Edge AB has the highest betweenness, because all paths between any circle and square must pass through it. If one were to remove it, the squares and circles would be split into two separate communities. The algorithm of Wilkinson et al. repeatedly identifies inter-community edges of large betweenness such as AB and removes them, until the graph is resolved into many separate communities.

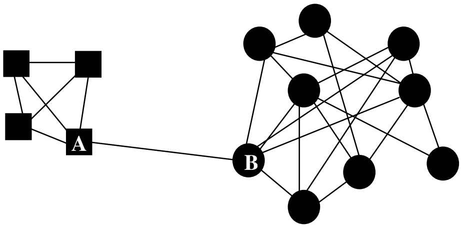

Fig. 1. An example graph with edge AB having high betweenness.

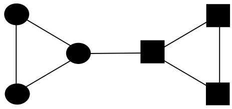

Fig. 2. The smallest possible graph of two viable communities.

Because the removal of an edge strongly affects the betweenness of many others, the values were repeatedly updated with the fast algorithm of Brandes [27, 28,24]. The procedure stops removing edges when it cannot further meaningfully subdivide communities. Figure 2 shows the smallest possible component that can be subdivided into two viable subcommunities. It has 6 nodes, consisting of two triangles linked by one edge. A component with fewer than 6 nodes cannot be subdivided further.

Components of size ≥ 6, for example the group of size nine in Fig. 1, can also constitute single cohesive communities. Figure 3 shows how the algorithm determines when to stop subdividing a community. The edge XY has the highest betweenness, but removing it would separate a single node, which does not constitute a viable community. In general, the single edge connecting a leaf vertex (such as X in Fig. 3) to the rest of a graph of N vertices has a betweenness of $N-1$, because it contains the shortest path from X to all $N-1$ other vertices. The stopping criterion for components of size ≥ 6 is therefore that the highest betweenness of any edge in the component be equal to or less than $N-1$.

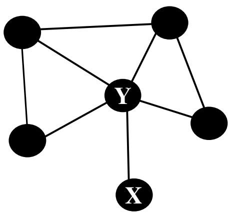

Fig. 3. An example graph of one community that does not contain distinct subcommunities.

2.2 Multiple Community Structures

As mentioned above, the removal of any one edge affects the betweenness of all the other edges, particularly in large, real-world graphs such as the email graph. Early in the process, there are many inter-community edges which have high betweenness and the choice of which to remove, while arbitrary, dictates which edges will be removed later. For example, a node belonging to two communities can be placed in one or the other by the algorithm, depending on the order in which edges are removed. One can take advantage of this arbitrariness to repeatedly partition the graph into many different "structures" or sets of communities. These sets are then compared and aggregated into a final list of communities.

Wilkinson and Huberman [25] introduced randomness into the algorithm by calculating the shortest paths from a random subset as opposed to all the nodes. The algorithm cycles randomly through at least m centers (where m is some cutoff) until the betweenness of at least one edge exceeds the threshold betweenness of a "leaf" vertex. The edge whose betweenness is highest at that point is removed, and the procedure is repeated until the graph has been separated into communities. The modified algorithm may occasionally remove an intra-community edge, but such errors are unimportant when a large number of structures is aggregated.

Applying this modified process n times yields n community structures imposed on the graph. One can then compare the different structures and identify communities. For example, after imposing 50 structures on a graph, one might find: a community of people A, B, C, and D in 25 of the 50 structures; a community of people A, B, C, D, and E in another 20; and one of people A, B, C, D, E and F in the remaining 5. This result is reported in the following way: A(50) B(50) C(50) D(50) E(25) F(5) which signifies that A, B, C, and D form a well-defined community, E is related to this community, but also to some other(s), and F is only slightly, possibly erroneously, related to it. For details of the aggregation procedure, please see [25].

The entire process of determining community structure within the graph is displayed below.

- For i iterations, repeat {
  1. Identify disjoint components of the graph.
  2. For each component, check to see if component is a community.
     - If so, remove it from the graph and output it.
     - If not, remove edges of highest betweenness, using the modified Brandes algorithm for large components, and the normal algorithm for small ones. Continue removing edges until the community splits in two.
  3. Repeat step 2 until all vertices have been removed from the graph in communities. }
- Aggregate the i structures into a final list of communities.

2.3 Results

The algorithm was applied to email data from the HP Labs mail server from the period November 25, 2002 to February 18, 2003, with 185,773 emails exchanged between the 485 HP Labs employees. For simplicity, emails that had an external origin or destination were omitted. Messages sent to a list of more than 10 recipients were likewise removed, as these emails were often lab-wide announcements (rather than personal communication), which were not useful in identifying communities of practice.

A graph was constructed from this data by placing edges between any two individuals that had exchanged at least 30 emails in total, and at least 5 in both directions. The threshold eliminated infrequent or one-way communication, and eliminated some individuals from the graph who either sent very few emails or used other email systems. The resulting graph consisted of 367 nodes, connected by 1110 edges.

There was one giant connected component of 343 nodes and six smaller components ranging in size from 2 to 8. The modified Brandes algorithm detected 60 additional distinct communities within the giant component. The largest community consisted of 57 individuals. The mean community size was 8.4, with standard deviation 5.3. A comparison of these communities with information from the HP corporate directory revealed that 49 of the 66 communities consisted of individuals entirely within one lab or organizational unit. The remaining 17 contained individuals from two or more organizations within the company.

2.4 Identifying Leadership Roles

In addition to identifying formal and informal work communities, it is also possible to draw inferences about the leadership of an organization from its communication data. One method is to visualize the above graph of the HP Labs email network with a standard force-directed spring algorithm [29], shown in Fig. 4. This spring layout of the email network does not use any information about the actual organization structure, and yet high level managers (the reddest nodes are at the top of the hierarchy) are placed close to the center of the graph. The trend is quantified in Table 1, which lists the average hierarchy depth (levels from the lab director) as a function of the position in the layout from the center.

Note that there is a group of 6 nodes in the upper right portion of the graph that are quite removed from the center, but are relatively high in the organizational hierarchy. This is the university relations group that reports directly to the head of HP Labs, but has no other groups reporting to it. Hence the layout algorithm correctly places them on the periphery of the graph, since their function, that of managing HP's relationship with universities, while important, is not at the core of day-to-day activities of the labs.

Evaluating communication networks with this technique could provide information about leadership in communities about which little is known. Sparrow proposed this approach for analyzing criminal networks [30], noting that "Euclidean centrality is probably the closest to the reality" of the current criminal

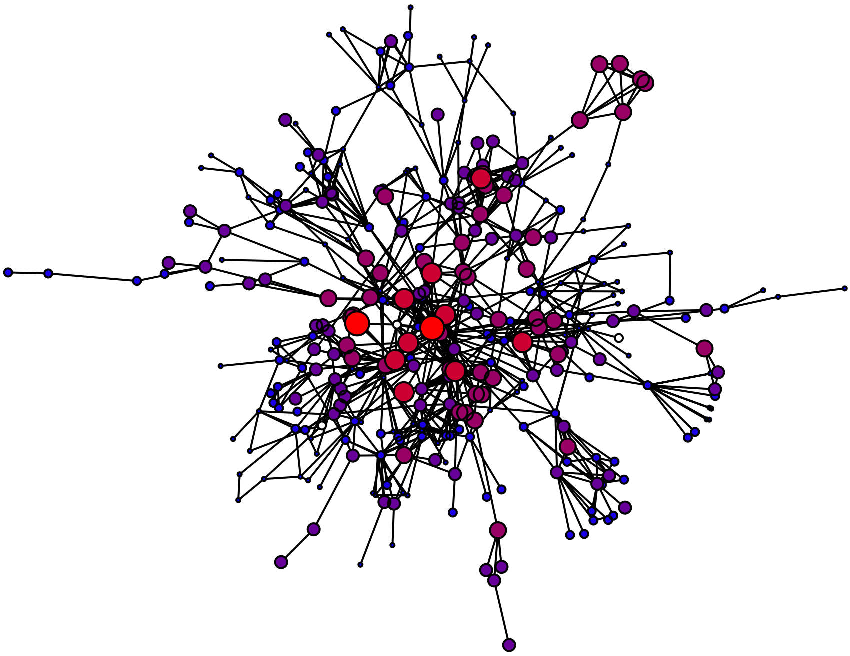

Fig. 4. The giant connected component of the HP Labs email network. The redness and size of a vertex indicates an individual's closeness to the top of the lab hierarchy (red-close to top, blue-far from top, black-no data available).

Table 1. Average hierarchy depth by distance from center in layout

distance from center	number of vertices	average depth in hierarchy
< 0.1	14	2.6
0.1 to 0.2	32	3.0
0.2 to 0.3	56	3.2
0.3 to 0.4	66	4.0
0.4 to 0.5	56	4.0
0.5 to 0.6	45	4.2
0.6 to 0.7	42	4.0
0.7 to 0.8	12	3.9
0.8 to 0.9	13	3.8

network analysis techniques. More recently, Krebs applied centrality measures and graphing techniques [31] to the terrorist networks uncovered in the 9/11 aftermath. He found that the average shortest path was unusually long for such a small network, and concluded that the operation had traded efficiency for secrecy - individuals in one part of the network did not know those in other parts of the network. If one cell had been compromised, the rest of the network would remain relatively unaffected. Several social network centrality measures pointed to Mohamed Atta's leadership role in the attacks of Sept. 11. The role was also confirmed by Osama bin Laden in a video tape following the attacks.

2.5 Field Evaluation

The HP Labs social network, being much less covert, could readily be compared to the structure of the formal organization. Nevertheless, the informal communities identified by the algorithm could not be verified in this way. Tyler et al. decided to validate the results of their algorithm by conducting a brief, informal field study. Sixteen individuals chosen from seven of the sixty communities identified were interviewed informally. The communities chosen represented various community sizes and levels of departmental homogeneity. They ranged in size from four to twelve people, and three out of the seven were heterogeneous (included members of at least two different departmental units within the company).

All sixteen subjects gave positive affirmation that the community reflected reality. More specifically, eleven described the group as reflecting their department, four described it as a specific project group, and one said it was a discussion group on a particular topic. Nine of the sixteen (56.25%) said nobody was missing from the group, six people (37.5%) said one person was missing, and one person (6.25%) said two people were missing. Conversely, ten of the sixteen (62.5%) said that everybody in the group deserved to be there, whereas the remaining six (37.5%) said that one person in the group was misclassified.

The interviews confirmed that most of the communities identified were based on organization structure. However, the communities also tended to include people who were de facto department members, but who did not technically appear in the department's organization chart, such as interns or people whose directory information had changed during the two months of the study. Finally, the algorithm seemed to succeed in dividing departmental groups whose work is distinct, but lumped together groups whose projects overlap.

Heterogeneous, cross-department communities are of particular interest because they cannot be deduced from the formal organization. The interviews revealed that most of them represented groups formed around specific projects, and in one case, a discussion forum. For example, one community contained three people from different labs coordinating on one project: a technology transfer project manager, a researcher who was the original designer of a piece of PC hardware, and an engineer redesigning the hardware for a specific printer.

2.6 Discussion

The power of this method for identifying communities and leadership is in its automation. It does an effective job of uncovering communities of practice with nothing more than email log ("to:" and "from:") data. The betweenness centrality measures can be further augmented to incorporate weights on the edges, representing, for example, the frequency of communication along a link [32].

Because the method of Wilkinson et al. [25] needs to re-run the Brandes algorithm every time an edge is removed, the algorithm has a running time of $O(n^3)$. Even faster algorithms, that can identify communities in $O(n^2)$ [33] and $O(n)$ [34] time have since been developed. The simplicity and speed of these new algorithms means that they can be applied to organizations of thousands to hundreds of thousands and produce results efficiently.

Communities identified in this automated way lack the richness in contextual description provided by ethnographic approaches. They do not reveal the nature or character of the identified communities, the relative importance of one community to another, or the subtle inter-personal dynamics within the communities. These kinds of details can only be uncovered with much more data- or labor-intensive techniques. However, in cases where an organization is very large, widely dispersed, or incompletely defined (informal), this method provides an suitable alternative or compliment to the more traditional, labor-intensive approaches.

3 Information Flow in Social Groups

In the previous section we saw that individuals tend to organize both formally and informally into groups based on their common activities and interests. In this section we examine how this structure in the interaction network affects the way information spreads. This is not unlike the transmission of an infectious agent among individuals, where the pattern of contacts determines how far a disease spreads. Thus one would expect that epidemic models on graphs are relevant to the study of information flow in organizations.

As we will show shortly (see, for example Fig. 8), e-mail networks can form scale free graphs. This is of particular interest, since 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 [35–37]. The presence of additional network structure was found to further influence the spread of disease on scale-free graphs [38–40].

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 [41–43]. Conversely, individuals many steps removed in a social network on average tend not to have as much in common, as shown in a study [19] of a network of Stanford student homepages and illustrated in Fig. 5.

Wu et al. [7] introduced 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. This epidemic model on a scale-free network has a finite threshold, implying that the spread

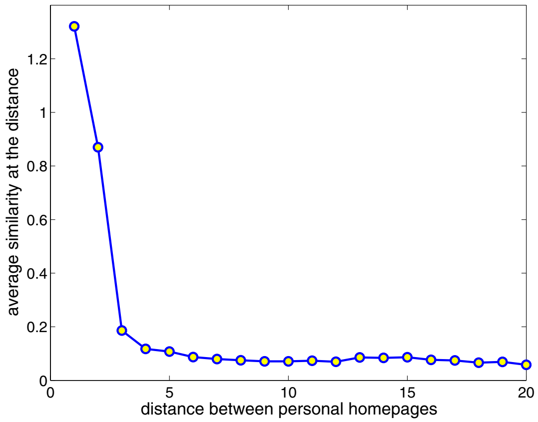

Fig. 5. Average similarity of Stanford student homepages as a function of the number of hyperlinks separating them.

of information is limited. The predictions were further tested 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 of interacting individuals, where the 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 a cutoff at the maximum degree $k = N$, where N is the number of nodes, and many networks show a cutoff at values much smaller than N. For their analysis, Wu et al. [7] made use of generating functions, whose application to graphs with arbitrary degree distributions is discussed in [44]. 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 [45] for the SIR (susceptible, infected, removed) model, one can 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 the information will be transmitted across an edge in the network from a infective individual to a susceptible neighbor. T is derived in [45] as a function of r_{ij}, the rate of contacts between the two nodes, and τ_i, the time a node remains infective. If we assume to a first approximation that 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 $p_m^{(1)}$ is given by

$$G^{(1)}(x) = \sum_{l=0}^{\infty} \sum_{k=l}^{\infty} p_k \binom{k}{l} T^l (1-T)^{k-l} x^l \tag{5}$$

$$= G_0(1 + (x-1)T) = G_0(x; T). \tag{6}$$

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 [44]. Similarly, if we define $G^{(m)}(x)$ to be 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

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

Or, more explicitly,

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

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

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

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

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

or

$$(m+1)^{-\beta}TG_1'(1) < 1. \qquad (16)$$

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

$$\langle s \rangle = \sum_{m=1}^{\infty} z_m \qquad (17)$$

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 6 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 [36]. Keeping β at zero and adding a cutoff at $\kappa = 100$ produces a non-zero critical transmissibility T_c, as was found in [45]. 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, a sample

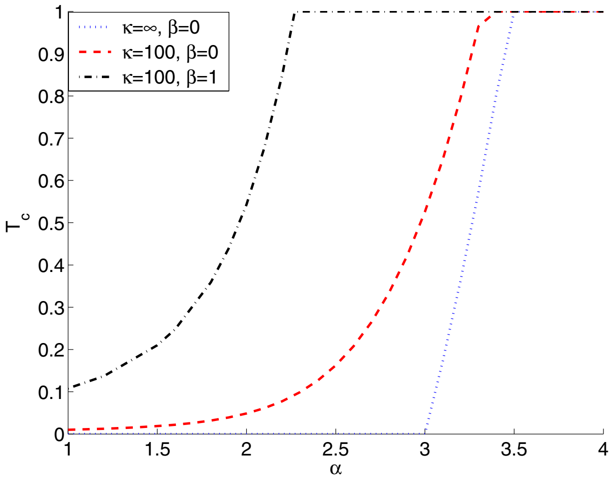

Fig. 6. T_c as a function of α. The three different curves, from bottom to top are: 1) no decay in transmission probability, no exponential cutoff in the degree distribution $(\kappa = \infty, \beta = 0)$. 2) $\kappa = 100, \beta = 0$, 3) $\kappa = 100, \beta = 1$.

from the mail clients of 40 individuals (30 within HP Labs, and 10 from other areas of HP, other research labs, and universities) was gathered. 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, the data was restricted 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 the sample was 2200, indicating that many users keep a substantial portion of their email correspondence. Although some messages may have been lost when users deleted them, it was assumed that a majority of messages containing useful information had been retained.

Figure 7 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 the study subjects, the content and significance of most of these messages was identified. 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.

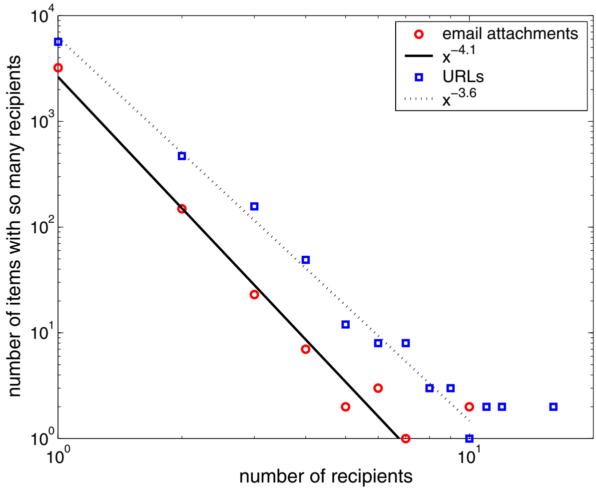

Fig. 7. Number of people receiving URLs and attachments

The next portion of the analysis analyzed the effect of decay in the transmission probability on the email graph at HP Labs. 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 Fig. 8, 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. A graph with the internal and external nodes mixed (as in [46]) was used to specifically demonstrate the effect of a decay on the spread of email in a power-law graph.

The spread of a piece of information was simulated 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 (one willing to transmit a particular piece of information) 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. It is also assumed that an individual remains infective for a period of 24 hours.

Next a decay was introduced 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. The exponent roughly corresponds to the decay in similarity between homepages shown in Fig. 5. 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 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

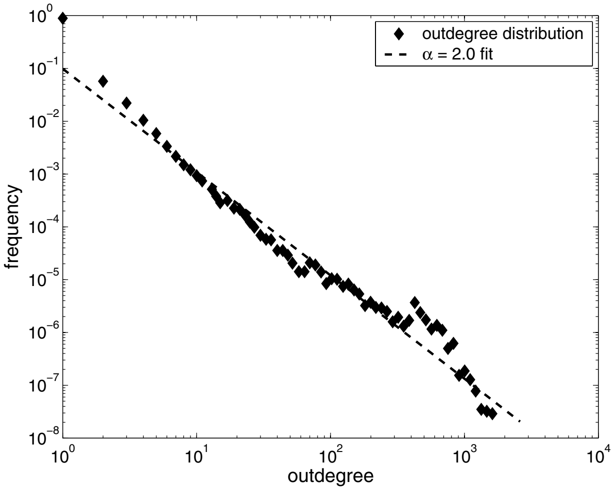

Fig. 8. Outdegree distribution for all senders (224,514 in total) sending email to or from the HP Labs email server over the course of 3 months. The outdegree of a node is the number of correspondents the node sent email to.

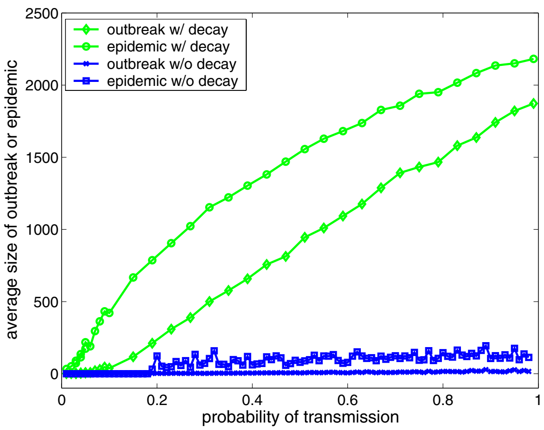

Fig. 9. Average outbreak and epidemic size as a function of the transmission probability p. The total number of potential recipients is 7119.

with. The distance between someone within HP labs and someone outside of HP labs was set to the maximum hierarchical distance of 8.

Figure 9 shows the variation in the average outbreak size, and the average epidemic size (chosen to be any outbreak affecting more than 30 individuals). 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.

4 Small World Search

In the preceding section we discussed how the tendency of like individuals to associate with one another can affect the flow of information within an organization. In this section we will show how one can take advantage of the very same network structure to navigate social ties and locate individuals.

The observation that any two people in the world are most likely linked by a short chain of acquaintances, known as the "small world" phenomenon has been the focus of much research over the last forty years [47–50]. In the 1960's and 70's, participants in small world experiments successfully found paths from Nebraska to Boston and from Los Angeles to New York. In an experiment in 2001 and 2002, 60,000 individuals were able to repeat the experiment using email to form chains with just four links on average across different contents [51]. The small world phenomenon is currently exploited by commercial networking services such as LinkedIn, Friendster, and Spoke[1] to help people network, for both business and social purposes.

The existence of short paths is not particularly surprising in and of itself. Although many social ties are "local" meaning that they are formed through one's work or place of residence, Watts and Strogatz [52] showed that it takes only a few "random" links between people of different professions or location to create short paths in a social network and make the world "small". In addition, Pool and Kochen [53] have estimated that an average person has between 500 and 1,500 acquaintances. Ignoring for the moment overlap in one's circle of friends, one would have $1,000^2$ or $1,000,000$ friends of friends, and $1,000^3$ or one billion friends-of-friends-of-friends. This means that it would take only 2 intermediaries to reach a number of people on the order of the population of the entire United States.

Although the existence of short paths is not surprising, it is another question altogether how people are able to select among hundreds of acquaintances the correct person to form the next link in the chain. Killworth and Barnard [50] performed the "reverse" experiment to measure how many acquaintances a typical person would use as a first step in a small world experiment. Presented with

[1] `http://www.linkedin.com/`,
`http://www.friendster.com/`,
`http://www.spokesoftware.com/`

1,267 random targets, the subjects chose about 210 different acquaintances on average, based overwhelmingly on geographic proximity and similarity of profession to the targets.

Recently, mathematical models have been proposed to explain why people are able to find short paths. The model of Watts, Dodds, and Newman [54] assumes that individuals belong to groups that are embedded hierarchically into larger groups. For example an individual might belong to a research lab, that is part of an academic department at a university, that is in a school consisting of several departments, that is part of a university, that is one of the academic institutions in the same country, etc. The probability that two individuals have a social tie to one another is proportional to $\exp^{-\alpha h}$, where h is the height of their lowest common branching point in the hierarchy.

The decay in linking probability means that two people in the same research laboratory are more likely to know one another than two people who are in different departments at a university. The model assumes a number of separate hierarchies corresponding to characteristics such as geographic location or profession. In reality, the hierarchies may be intertwined, for example professors at a university living within a short distance of the university campus, but for simplicity, the model treats them separately.

In numerical experiments, artificial social networks were constructed and a simple greedy algorithm was performed where the next step in the chain was selected to be the neighbor of the current node with the smallest distance along any dimension. At each step in the chain there is a fixed probability, called the attrition rate, that the node will not pass the message further. The numerical results showed that for a range of the parameter α and number of attribute dimensions, the networks are "searchable", meaning that a minimum fraction of search paths find their target.

Kleinberg [55,56] posed a related question: in the absence of attrition, when does the length of the chains scale in the same way as the average shortest path. Unlike the study of Watts. et al., there is no attrition - all chains run until completion, but need to scale as the actual shortest path in the network does. In the case of a small world network, the average shortest path scales as $\ln(N)$, where N is the number of nodes. Kleinberg proved that a simple greedy strategy based on geography could achieve chain lengths bounded by $(\ln N)^2$ under the following conditions: nodes are situated on an m-dimensional lattice with connections to their $2m$ closest neighbors and additional connections are placed between any two nodes with probability $p \sim r^{-m}$, where r is the distance between them. Since in the real world our locations are specified primarily by two dimensions, longitude and latitude, the probability is inversely proportional to the square of the distance. A person should be four times as likely to know someone living a block away, than someone two city blocks away. However, Kleinberg also proved that if the probabilities of acquaintance do not follow this relationship, nodes would not be able to use a simple greedy strategy to find the target in polylogarithmic time.

The models of both Watts et al. and Kleinberg show that the probability of acquaintance needs to be related to the proximity between individuals' attributes in order for simple search strategies using only local information to be effective. Below we describe experiments empirically testing the assumptions and predictions of the proposed two models.

4.1 Method

In order to test the above hypothesis, Adamic and Adar [8] applied search algorithms to email networks derived from the email logs at HP Labs already described in Sect. 2. A social contact was defined to be someone with whom an individual had exchanged at least 6 emails each way over the period of approximately 3 months. The bidirectionality of the email correspondence guaranteed that a conversation had gone on between the two individuals and hence that they are familiar with one another.

Imposing this constraint yielded a network of 436 individuals with a median number of 10 acquaintances and a mean of 13. The degree distribution, shown in Fig. 10, is highly skewed with an exponential tail. This is in contrast to the raw power-law email degree distribution, used in Sect. 3 and shown in Fig. 8, pertaining to both internal and external nodes and possessing no threshold in email volume. A scale free distribution in the raw network arises because there are many external nodes emailing just one individual inside the organization, and there are also some individuals inside the organization sending out announcements to many people and hence having a very high degree. However, once we impose a higher cost for maintaining a social contact (that is, emailing that contact at least six times and receiving at least as many replies), then there are few individuals with many contacts.

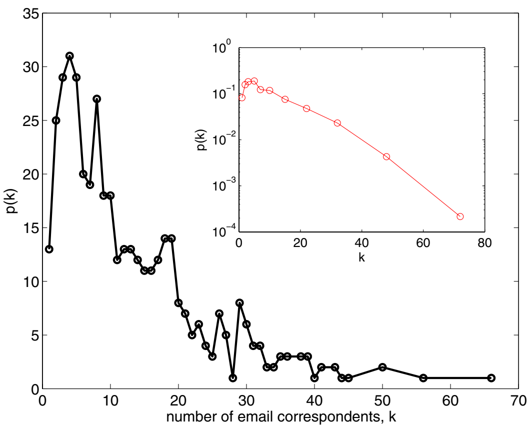

Fig. 10. Degree distribution in the HP Labs email network. Two individuals are linked if they exchanged at least 6 emails in either direction. The inset shows the same distribution, but on a semilog scale, to illustrate the exponential tail of the distribution

4.2 Simulating Milgram's Experiment on an Email Network

The resulting network, consisting of regular email patterns between HP Labs employees, had 3.1 edges separating any two individuals on average, and a median of 3. Simulations were performed on the network to determine whether members of the network would be able to use a simple greedy algorithm to locate a target. In this simple algorithm, each individual can use knowledge only of their own email contacts, but not their contacts' contacts, to forward the message.

Three different strategies were tested, at each step passing the message to the contact who is either

- best connected
- closest to the target in the organizational hierarchy
- sitting in closest physical proximity to the target

The first strategy selects the individual who is more likely to know the target by virtue of the fact that he/she knows so many people. It has been shown [57], that this is an effective strategy in power-law networks with exponents close to 2 (the case of the unfiltered HP Labs email network), but that it performs poorly in graphs with a Poisson degree distribution that has an exponential tail. Since the distribution of contacts in the filtered HP network was not power-law, the high degree strategy was not expected to perform well, and this was verified through simulation. The median number of steps required to find a randomly chosen target from a random starting point was 17, compared to the three steps in the average shortest path. Even worse, the average number of steps is 40. This discrepancy between the mean and the median is a reflection of the skewness of the distribution: a few well connected individuals and their contacts are easy to find, but some individuals who do not have many links and are not connected to highly connected individuals are difficult to locate using this strategy.

The second strategy consisted of passing the message to the contact closest to the target in the organizational hierarchy. The strategy relies on the observation, illustrated in Figs. 11 and 13 that individuals closer together in the organizational hierarchy are more likely to email with one another. Figure 12 illustrates such a search, labelling nodes by their hierarchical distance (h-distance) from the target. The h-distance is computed as follows: a node has distance one to their manager and to everyone they share a manager with. Distances are then recursively assigned, so that each node has h-distance 2 to their first neighbor's neighbors, and h-distance 3 to their second neighbor's neighbors, etc. A simple greedy strategy using information about the organizational hierarchy worked extremely well. The median number of steps was only 4, close to the median shortest path of 3. With the exception of one individual, whose manager was not located on site, and who was consequently difficult to locate, the mean number of steps was 4.7, meaning that not only are people typically easy to find, but nearly everybody can be found in a reasonable number of steps.

In the original experiment by Milgram the completed chains were divided between those that reached the target through his professional contacts and those that reached him through his hometown. On average those that relied on

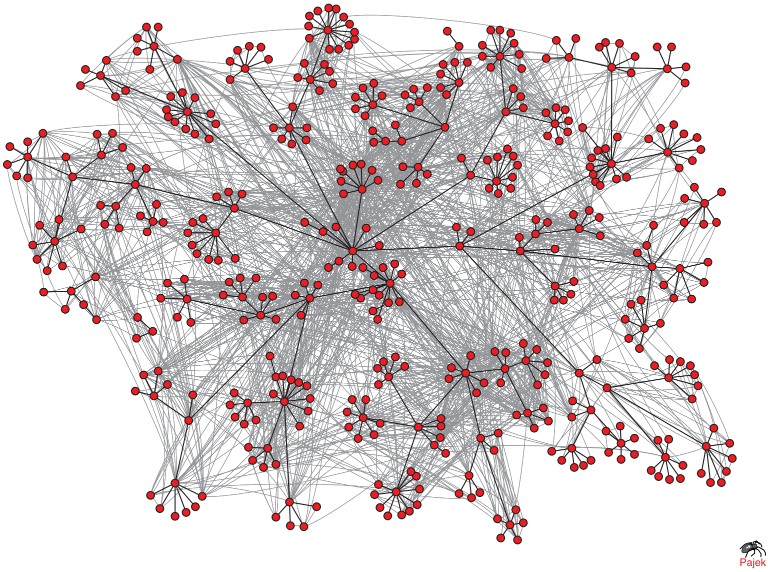

Fig. 11. Email communications within HP Labs (gray lines) mapped onto the organizational hierarchy (black lines). Note that email communication tends to "cling" to the formal organizational chart.

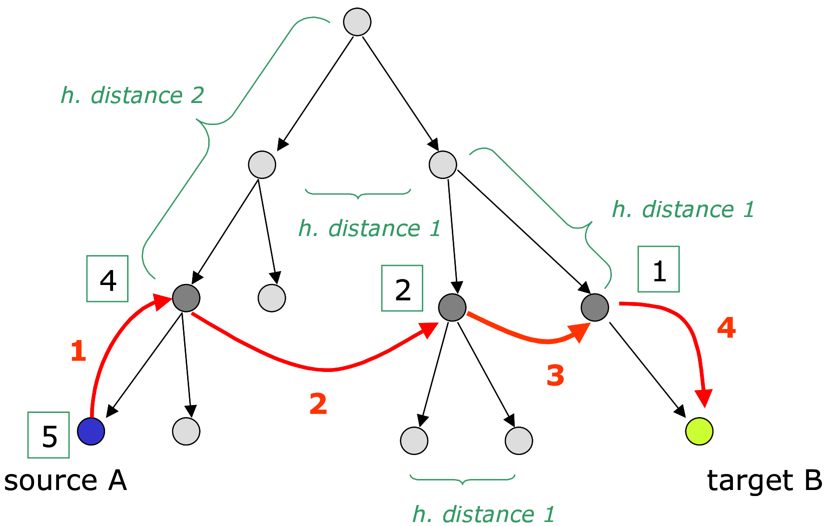

Fig. 12. Example illustrating a search path using information about the target's position in the organizational hierarchy to direct a message. Numbers in the square give the h-distance from the target.

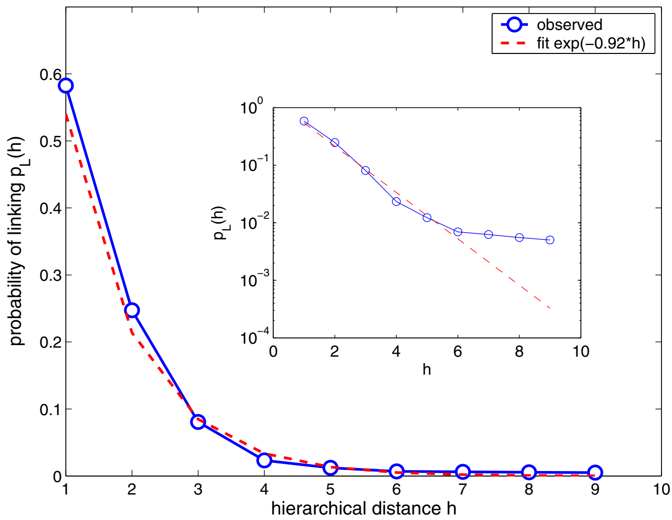

Fig. 13. Probability of linking as a function of the separation in the organizational hierarchy. The exponential parameter $\alpha = 0.92$, in the searchable range according to the model of Watts et al.[54]

geography took 1.5 steps longer to reach the target, a difference found to be statistically significant. In the words of Travers and Milgram [48], the following seemed to occur: "Chains which converge on the target principally by using geographic information reach his hometown or the surrounding areas readily, but once there often circulate before entering the target's circle of acquaintances. There is no available information to narrow the field of potential contacts which an individual might have within the town."

Performing the small world experiment on the HP email network using geography produced a similar result, in that geography could be used to find most individuals, but was slower, taking a median number of 7 steps, and a mean of 12. Figure 14 shows the email correspondence mapped onto the physical layout of the buildings. Individuals' locations are given by their building, the floor of the building, and the nearest building post (for example "H15") to their cubicle. The distance between two cubicles was approximated by the "street" distance between their posts (for example "A3" and "C10" would be $(C - A) \times 25' + (10 - 3) \times 25' = 2 \times 25' + 7 \times 25' = 225$ feet apart). Adding the x and y directions separately reflects the interior topology of the buildings where one navigates perpendicular hallways and cannot traverse diagonally. If individuals are located on different floors or in different buildings, the distance between buildings and the length of the stairway are factored in.

The general tendency of individuals in close physical proximity to correspond holds: over 87% percent of the 4000 email links are between individuals on the same floor, and overall individuals closer together are more likely to correspond. Still, individuals maintain disproportionately many far-flung contacts while not

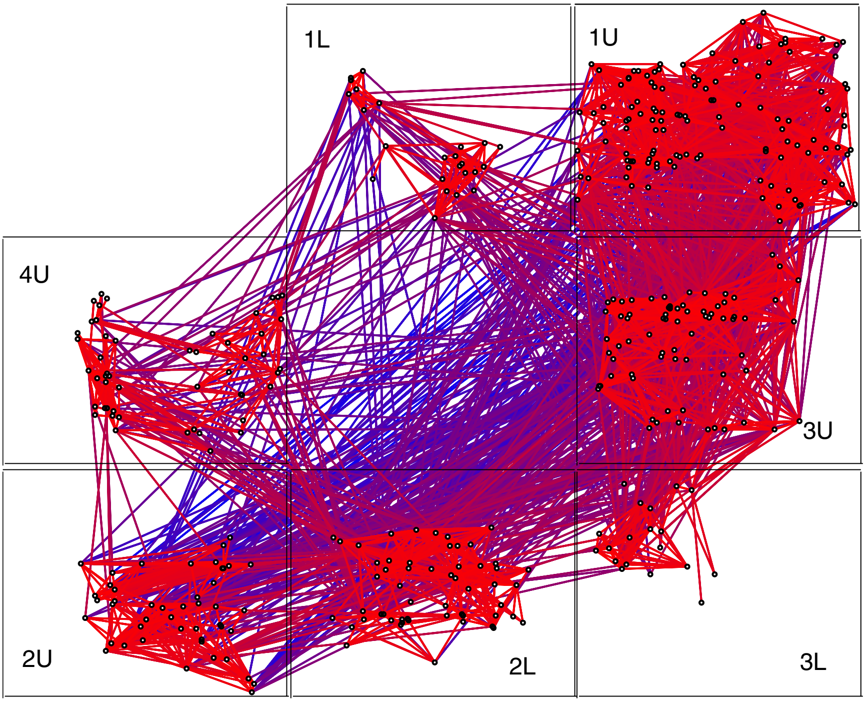

Fig. 14. Email communications within HP Labs mapped onto approximate physical location based on the nearest post number and building given for each employee. Each box represents a different floor in a building. The lines are color coded based on the physical distance between the correspondents: red for nearby individuals, blue for far away contacts.

getting to know some of their close-by neighbors. The relationship between probability of acquaintance and cubicle distance r between two individuals, shown in Fig. 15, is well-fitted by a $1/r$ curve. However, Kleinberg has shown that the optimum relationship in two dimensional space is $1/r^2$ - a stronger decay in probability of acquaintance than the $1/r$ observed.

In the case of HP Labs, the geometry may not be quite two dimensional, because it is complicated by the particular layout of the buildings. Hence the optimum relationship may lie between $1/r$ and $1/r^2$. In any case, the observed $1/r$ probability of linking shows a tendency consistent with Milgram's observations about the original small world experiment. At HP Labs, because of space constraints, re-organizations, and personal preferences, employees' cubicles may be removed from some of the co-workers they interact with. This hinders a search strategy relying solely on geography, because one might get physically quite close to the target, but still need a number of steps to find an individual who interacts with them.

Figure 16 shows a histogram of chain lengths resulting from searches using each of the three strategies. It shows the clear advantage of using the target's

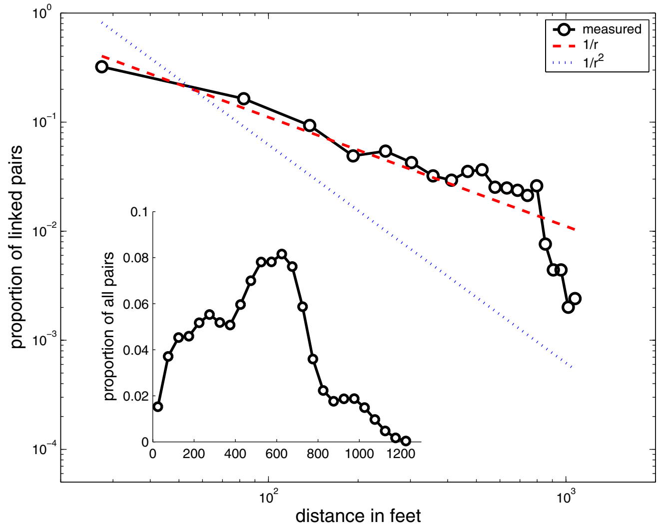

Fig. 15. Probability of two individuals corresponding by email as a function of the distance between their cubicles. The inset shows how many people in total sit at a given distance from one another.

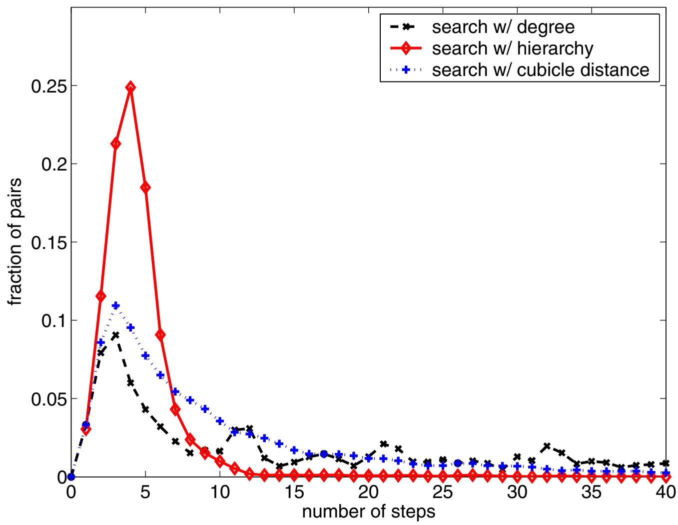

Fig. 16. Results of search experiments utilizing either knowledge of the target's position in the organizational hierarchy or the physical location of their cubicle.

position in organizational hierarchy as opposed to his/her cubicle location to pass a message through one's email contact. It also shows that both searches using information about the target outperform a search relying solely on the connectivity of one's contacts.

4.3 Discussion

The above simulated experiments verify the models proposed in [54] and [55] to explain why individuals are able to successfully complete chains in the small world experiments using only local information. When individuals belong to groups based on a hierarchy and are more likely to interact with individuals within the same small group, then one can safely adopt a greedy strategy - pass the message onto the individual most like the target, and they will be more likely to know the target or someone closer to them.

At the same time it is important to note that the optimum relationship between the probability of acquaintance and distance in physical or hierarchical space between two individuals, as outlined in [55,56], are not exactly satisfied. We just saw that the relationship between the physical distance and the probability of corresponding by email follows an inverse rather than an inverse square relationship. There are too many distant contacts and too few nearby ones compared to the optimum. A similar, albeit weaker trend holds for organizational distance. In Section 2 email spectroscopy revealed that while collaborations mostly occurred within the same organizational unit, they also frequently bridged different parts of the organization or broke up a single organizational unit into noninteracting subgroups. The optimum relationship derived in [56] for the probability of linking would be inversely proportional to the size of the smallest organizational group that both individuals belong to. However, the observed relationship, shown in Fig. 17 is slightly off, with $p \sim g^{-3/4}$, g being the group size.

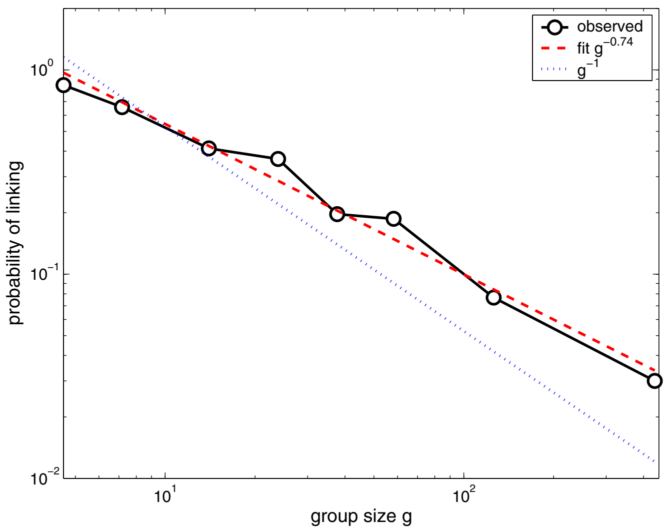

Fig. 17. Probability of two individuals corresponding by email as a function of the size of the smallest organizational unit they both belong to. The optimum relationship derived in [56] is $p \sim g^{-1}$, g being the group size. The observed relationship is $p \sim g^{-3/4}$.

Overall, the results of the email study are consistent with the model of Watts et al. [54]. This model does not require the search to find near optimum paths, but simply determines when a network is "searchable", meaning that fraction of messages reach the target given a rate of attrition. The relationship found between separation in the hierarchy and probability of correspondence, shown in Fig. 13, is well within the searchable regime identified in the model.

The study of Adamic and Adar is a first step, validating these models on a small scale. The email study gives a concrete way of observing how the small world chains can be constructed. Using a very simple greedy strategy, individuals across an organization could reach each other through a short chain of coworkers. It is quite likely that similar relationships between acquaintance and proximity (geographical or professional) hold true in general, and therefore that small world experiments succeed on a grander scale for the very same reasons.

5 Conclusion

In this chapter we reviewed three studies of information flow in social networks. The first developed a method of analyzing email communication automatically to expose communities of practice and their leaders. The second showed that the tendency of individuals to associate according to common interests influences the way that information spreads throughout a social group. It spreads quickly among individuals to whom it is relevant, but unlike a virus, is unable to infect a population indiscriminately. The third study showed why small world experiments work - how individuals are able to take advantage of the structure of social networks to find short chains of acquaintances. All three studies relied on email communication to expose the underlying social structure, which previously may have been difficult and labor-intensive to obtain. We expect that these findings are also valid with other means of social communication, such as verbal exchanges, telephony and instant messenger systems.

Acknowledgements

We would like to thank Eytan Adar, Joshua Tyler, and TJ Giuli for their comments and suggestions.

References

1. B. Wellman, "Computer networks as social networks," *Science*, vol. 293, pp. 2031–34, 2002.
2. S. Whittaker and C. Sidner, "Email overload: exploring personal information management of email," in *Proceedings of CHI'96 Conference on Computer Human Interaction*, (New York), pp. 276–283, Logos Verlag, 21996.
3. R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in organizations," *Physical Review E*, vol. 65, p. 065103, 2003.

4. J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as spectroscopy: Automated discovery of community structure within organizations," in *Proceedings of the International Conference on Communities and Technologies*, (Netherlands), Kluwer Academic Publishers, 2003.
5. J.-P. Eckmann, E. Moses, and D. Sergi, "Dialog in e-mail traffic." `http://xyz.lanl.gov/abs/cond-mat/0304433"`, 2003.
6. M. F. Schwartz and D. C. M. Wood, "Discovering shared interests among people using graph analysis," *Communications of the ACM*, vol. 36, no. 8, pp. 78–89, 1993.
7. F. Wu, B. A. Huberman, L. A. Adamic, and J. Tyler, "Information flow in social groups," *Physica A*, 2004. (to appear).
8. L. A. Adamic and E. Adar, "How to search a social network." submitted for publication, `http://www.hpl.hp.com/shl/papers/socsearch`, 2002.
9. W. G. Ouchi, "Markets, bureaucracies, and clans," *Administrative Science Quarterly*, vol. 25, pp. 129–141, 1980.
10. P. Blau and W. Scott, *Formal organizations. A comparative approach.* Lndn. Routledge & Kegan Paul, 1963.
11. R. S. Burt, "Models of network structure," *Annual Review of Sociology*, vol. 6, pp. 79–141, 1980.
12. S. Wasserman and K. Faust, *Social network analysis.* Cambridge: Cambridge University Press, 1994.
13. W. R. Scott, *Organizations: Rational, Natural, and Open Systems.* Englewood Cliffs, NJ: Prentice-Hall, 1992.
14. B. Huberman and T. Hogg, "Communities of practice: Performance and evolution," *Computational and Mathematical Organization Theory*, vol. 1, pp. 73–92, 1995.
15. M. Crozier, *The Bureaucratic Phenomenon.* Chicago: University of Chicago Press, 1964.
16. D. Crane, *Invisible Colleges: Diffusion of Knowledge in Scientific Communities.* Chicago: University of Chicago Press, 1972.
17. J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press, 1991.
18. N. Ducheneaut and V. Bellotti, "A study of email work processes in three organizations." to appear in the Journal of CSCW.
19. L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.
20. T. Allen, *Managing the Flow of Technology.* Cambrige, MA: MIT Press, 1984.
21. P. Hinds and S. Kiesler, "Communication across boundaries: Work, structure, and use of communication technologies in a large organization," *Organization Science*, vol. 6, no. 4, pp. 373–393, 1995.
22. L. Garton, C. Haythornwaite, and B. Wellman, "Studying on-line social networks," *Journal of Computer Mediated Communication*, vol. 3, no. 1, 1997.
23. H. Alani, K. O'Hara, and N. Shadbolt, "Ontocopi: Methods and tools for identifying communities of practice, intelligent information processing conference," in *IFIP World Computer Congress (WCC)*, 2002.
24. M. Girvan and M. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 8271–8276, 2002.
25. D. Wilkinson and B. A. Huberman, "A method for finding communities of related genes." *Proc. Natl. Acad. Sci. USA*, Feb. 2, 2004, 10.1073/pnas.0307740100.
26. L. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

27. U. Brandes, “A faster algorithm for betweenness centrality,” *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.
28. M. Newman, “Who is the best connected scientist? a study of scientific coauthorship networks,” *Phys. Rev. E*, vol. 64, p. 016131, 2001.
29. T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Software - Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.
30. M. K. Sparrow, “The application of network analysis to criminal intelligence: An assessment of the prospects,” *Social Networks*, vol. 13, pp. 251–274, 1991.
31. V. E. Krebs, “Uncloaking terrorist networks,” *First Monday*, vol. 7, April 2002.
32. M. E. J. Newman, “Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality,” *Phys. Rev. E*, vol. 64, p. 016132, 2001.
33. M. E. J. Newman, “Fast algorithm for detecting community structure in networks.” `http://aps.arxiv.org/abs/cond-mat/0309508/`, 2003.
34. F. Wu and B. A. Huberman, “Discovering communities in linear time: a physics approach,” *Europhysics Letters*, 2004. to appear.
35. Z. Dezso and A.-L. Barabasi, “Halting viruses in scale-free networks,” *Phys. Rev. E*, vol. 65, p. 055103, 2002.
36. R. Pastor-Satorras and A. Vespignani, “Epidemic spreading in scale-free networks,” *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, 2001.
37. M. E. J. Newman, S. F., and J. Balthrop, “Email networks and the spread of computer viruses,” *Phys. Rev. E*, vol. 66, p. 035101, 2002.
38. V. M. Eguiluz and K. Klemm, “Epidemic threshold in structured scale-free networks,” *Phys. Rev. Lett.*, vol. 89, p. 108701, 2002.
39. A. Vazquez, M. Boguna, Y. Moreno, R. Pastor-Satorras, and A. Vespignani, “Topology and correlations in structured scale-free networks,” *Physical Review E*, vol. 67, p. 046111, 2003.
40. M. E. J. Newman, “Assortative mixing in networks,” *Phys. Rev. Lett*, vol. 89, p. 208701, 2002.
41. P. Lazarsfeld and R.K.Merton, “Friendship as a social process: A substantive and methodological analysis,” in *Freedom and Control in Modern Society* (M. Berger, T. Abel, and C. Page, eds.), New York: Van Nostrand, 1954.
42. J. Touhey, “Situated identities, attitude similarity, and interpersonal attraction,” *Sociometry*, vol. 37, pp. 363–374, 1974.
43. S. Feld, “The focused organization of social ties,” *American Journal of Sociology*, vol. 86, pp. 1015–1035, 1981.
44. M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distribution and their applications,” *Phys. Rev. E*, vol. 64, p. 026118, 2001.
45. M. Newman, “The spread of epidemic disease on networks,” *Phys. Rev. E*, vol. 66, p. 016128, 2002.
46. H. Ebel, L.-I. Mielsch, and S. Bornholdt, “Scale-free topology of e-mail networks,” *Phys. Rev. E*, vol. 66, p. 035103, 2002.
47. S. Milgram, “The small-world problem,” *Psychology Today*, vol. 1, pp. 62–67, 1967.
48. J. Traver and S. Milgram, “An experimental study of the small world problem,” *Sociometry*, vol. 32, pp. 425–443, 1969.
49. C. C. Lundberg, “Patterns of acquaintanceship in society and complex organization: A comparative study of the small world problem,” *Pacific Sociological Review*, vol. 18, pp. 206–222, 1975.
50. P. Killworth and H. Bernard, “Reverse small world experiment,” *Social Networks*, vol. 1, pp. 159–192, 1978.
51. P. S. Dodds, R. M., and D. J. Watts, “An experimental study of search in global social networks,” *Science*, vol. 301, pp. 827–829, 2003.

52. D. J. Watts and S. H. Strogatz, “Collective dynamics of small-world networks,” *Nature*, vol. 393, pp. 440–442, 1998.
53. I. Pool and M. Kochen, “Contacts and influence,” *Social Networks*, vol. 1, pp. 5–51, 1978.
54. D. J. Watts, P. S. Dodds, and M. E. J. Newman, “Identity and search in social networks,” *Science*, vol. 296, pp. 1302–1305, 2002.
55. J. Kleinberg, “Navigation in a small world,” *Nature*, vol. 406, 2000.
56. J. Kleinberg, “Small-world phenomena and the dynamics of information,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, 2001.
57. L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, “Search in power-law networks,” *Phys. Rev. E*, vol. 64, p. 046135, 2001.