How to search a social network

Lada A. Adamic
ladamic@hpl.hp.com

Eytan Adar
eytan@hpl.hp.com

HP Labs, 1501 Page Mill Road
Palo Alto, CA 94301

Abstract

We address the question of how participants in a small world experiment are able to find short paths in a social network using only local information about their immediate contacts. We simulate such experiments on a network of actual email contacts within an organization as well as on a student social networking website. On the e-mail network we find that small world search strategies using a contact's position in physical space or in an organizational hierarchy relative to the target can effectively be used to locate most individuals. However, we find that in the online student network, where the data is incomplete and hierarchical structures are not well defined, local search strategies are less effective. We compare our findings to recent theoretical hypotheses about underlying social structure that would enable these simple search strategies to succeed and discuss the implications to social software design.

1 Introduction

Many tasks, ranging from collaboration within and between organizations, pursuit of hobbies, or forming romantic relationships, depend on finding the right people to partner with. Sometimes it is advantageous to seek an introduction through one's contacts, who could recommend one to the desired target. Finding such paths through a network of acquaintances is something people naturally do, for example, when looking for a job. How people are actually able to this, while using only local information about the network, is a problem we address by analyzing real-world social networks. This analysis examines whether social networks are structured in a way to allow effective local search. In answering this question we have obtained insights that may be applicable to new commercial services, such as LinkedIn, Friendster, and Spoke[1], that have recently sprung up to help people network.

Social networking services gather information on users' social contacts, construct a large interconnected social network, and reveal to users how they are connected to others in the network. The premise of these businesses is that individuals might be

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

only a few steps removed from a desirable business or social partner, but not realize it. The services allow their users to get to know one's friends of friends and hence expand their own social circle. On a smaller scale, the Club Nexus online community, which we describe later on in this paper, sought to help students at Stanford University organize activities and find others with common interests through their social network.

Although the online social networking trend may be fairly recent, 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, is not. It has been the focus of much research over the last forty years [5, 8, 9, 11]. In the 1960's and 70's, participants in small world experiments successfully found paths connecting individuals from Nebraska to Boston and from Los Angeles to New York. In 2002, 60,000 individuals were able to repeat the experiment using email chains with an average of 4.1 links to bridge continents [3].

The existence of short paths, which is the essence of the small world phenomenon, is not particularly surprising in and of itself. This is because in the case of random acquaintance networks, the number of people grows exponentially with distance in the social network. If we take the average number of acquaintances to be about 1000 (Pool and Kochen [10] estimated in 1978 that the number lies between 500 and 1,500), 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.

The above calculation assumed that the network is random. That is, the overlap in one's friends and one's friend's friends is negligible. In reality, social networks are far from random. Most of one's contacts are formed through one's place of residence and profession, forming tightly knit cliques. This means many of one's friends' friends already belong to the set of one's own friends. Still, as was shown by Watts and Strogatz [14] 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".

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 [5] have 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 1,267 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 [13] 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, 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, and α is the decay parameter.

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

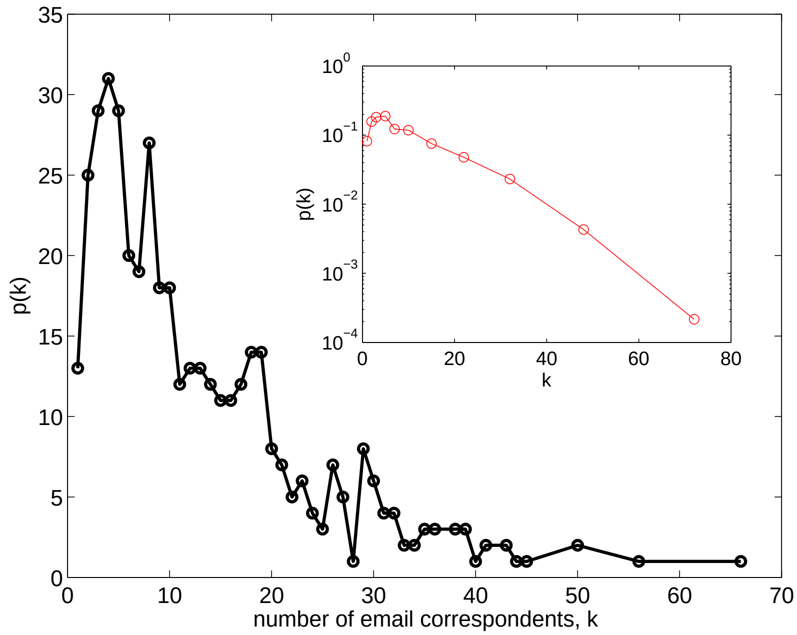

Figure 1: Degree distribution in the HP Labs email network. Two individuals are linked if they exchanged at least 6 emails in either direction. The inset shows the same distribution, but on a semilog scale, to illustrate the exponential tail of the distribution

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 on artificial social networks, a simple greedy search algorithm selected the next step in the chain to be the neighbor of the current node who is most similar to the target along any dimension. Each node in the chain has a fixed probability, called the attrition rate, of not passing the message further. The results of the numerical experiments showed that for a range of the decay parameter α and a number of attribute dimensions, the constructed networks are "searchable", where a "searchable" network is defined to be one where a minimum fraction of search paths find their target before attrition terminates them.

Kleinberg [6, 7] posed a related question: how social networks need to be structured in order for a simple greedy strategy to find near optimal paths through the network. 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 $2 * m$ closest neighbors and additional connections are placed between any two nodes with probability $p \sim r^{-m}$, where r is the Euclidean distance between them. Since individuals locations in the real world are specified primarily by two dimensions, longitude and latitude, the probability p should be inversely proportional to the square of the distance. A person would 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. Kleinberg also derived results for individuals belonging to hierarchically nested groups. If the probability of two people linking to one another is inversely proportional to the size of the smallest group that they both belong to, then greedy search can be used to find short paths 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.

1.1 Paper Roadmap

In this paper we examine the efficiency of social search simulated on two data sets. The first, a network based on e-mail communication will be discussed in Section 2. In this case, social relationships were inferred from email exchanges and represent a fairly complete picture of the communication network. In Section 3 we will describe search on a network extracted from a social networking website, containing partial information about the true network. We conclude by discussing how the differences between networks constructed through e-mail analysis and the website impact search performance.

2 Search in an E-Mail Network

We first used a social network derived from the email logs at HP Labs to test the assumptions of the theoretical models regarding the structure of social networks. We then tested whether simple greedy strategies can efficiently find short paths when the assumptions are satisfied. From the email logs, we defined a social contact to be someone with whom an individual had exchanged at least 6 emails both ways over the period of approximately 3 months. The bidirectionality of the email correspondence guaranteed that a conversation had taken place between the two individuals and hence that the two individuals know one another.

Imposing a communication threshold yielded a network of 436 individuals with a median number of 10 acquaintances and a mean of 13. The degree distribution, shown in Figure 1, is highly skewed with an exponential tail. The resulting network, consisting of regular email patterns between HP Labs employees, had 3.1 hops separating any two individuals on average, and a median of 3.

In the simulated search experiments, we considered three different properties of the nodes: degree, position in the organizational hierarchy, and physical location. In this simple algorithm, each individual can use knowledge only of their own email contacts, but not their contacts' contacts, to forward the message. We tested three corresponding strategies, at each step passing the message to the contact who is

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

The first strategy is a high-degree seeking strategy and 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 [2], that high degree seeking strategies are effective in networks with a power-law degree distribution with an exponent γ close to 2. In a power-law network, the probability of having k contacts is $p(k) \sim k^{-\gamma}$. This is precisely the degree distribution of the unfiltered HP Labs email network, where all communication, including unidirectional and infrequent correspondence, is taken into account [15]. The power-law distribution in the raw network arises because there are many external nodes emailing just a few individuals 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, one would be unlikely to utilize these one-time connections when searching for an individual for a task such as a job search, where the strength of the tie is important.

Once we impose a threshold (emailing at least six times and receiving at least as many replies) to identify a true social contact, fewer individuals have a high degree. As we showed above, the filtered network does not have a power-law degree distribution, but rather an exponential tail, similar to a Poisson distribution. Adamic et al.[2] showed that a search strategy attempting to use high degree nodes in a Poisson network performs poorly.

Simulation confirmed that the high degree seeking search strategy was unsuitable for the filtered HP email network. 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 was 40. This discrepancy between the mean and median is a reflection of the skewness of the distribution: a few well connected individuals and their contacts are easy to find, but others who do not have many links and are not connected to highly connected individuals are difficult to locate using this strategy.

The second strategy consists of passing the message to the contact closest to the target in the organizational hierarchy. In our simulation individuals are allowed full knowledge of the organizational hierarchy (in actuality, employees can reference an online organizational chart). However, the communication network that they are trying to navigate is hidden to them beyond their immediate contacts. The search strategy relies on the observation, illustrated in Figures 2 and 4, that individuals closer together in the organizational hierarchy are more likely to email one another.

Figure 3 illustrates such a search, labelling nodes by their hierarchical distance (h-distance) from the target. At each step in the chain the message is passed to someone closer in the organizational hierarchy to the target. Note that the message does not need to travel all the way to the top of the organizational hierarchy and instead takes advantage of a shortcut created by individuals in different groups communicating

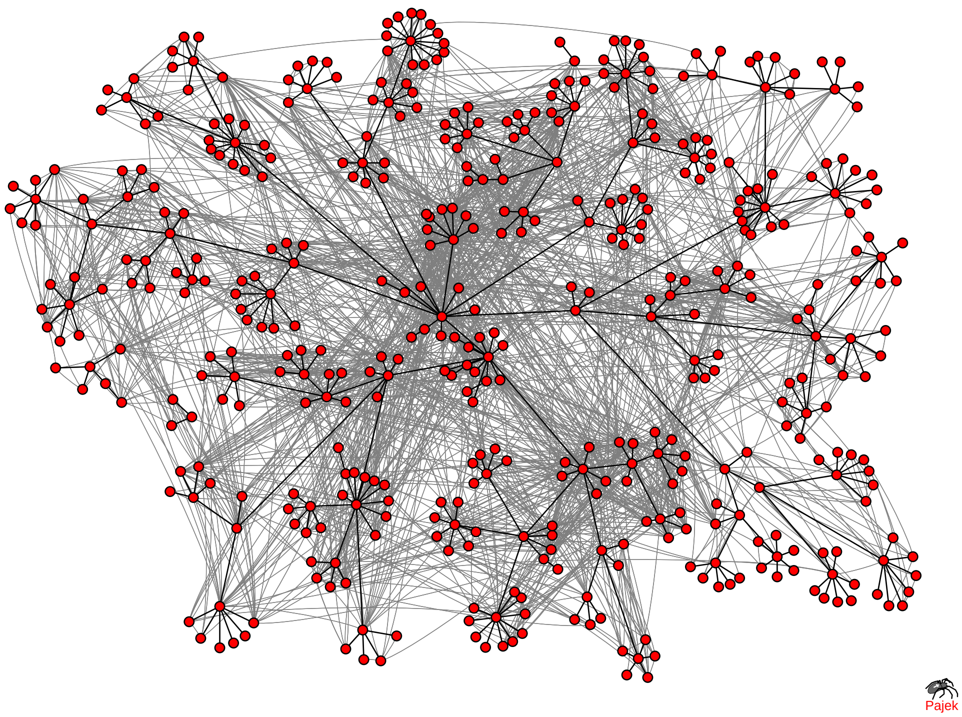

Figure 2: HP Labs' email communication (light grey lines) mapped onto the organizational hierarchy (black lines). Note that communication tends to "cling" to the formal organizational chart.

with one another. The h-distance, used to navigate the network, is computed as follows: individuals have h-distance one to their manager and to everyone they share a manager with. Distances are then recursively assigned, so that each individual has h-distance 2 to their first neighbor's neighbors, and h-distance 3 to their second neighbor's neighbors, etc.

The optimum relationship derived in [7] 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 Figure 5 is slightly off, with $p \sim g^{-3/4}$, g being the group size. This means that far-flung collaborations occur slightly more often than would be optimal for the particular task of searching, at the expense of short range contacts. The tendency for communication to occur across the organization was also revealed in an analysis utilizing spectroscopy methods on the same email network [12]. While collaborations mostly occurred within the same organizational unit, they also occasionally bridged different parts of the organization or broke up a single organizational unit into noninteracting subgroups.

Given the close correspondence between the assumptions of the models regarding group structure and the email network, we expected greedy strategies using the organizational hierarchy to work fairly well. Indeed, this was confirmed in our simulations.

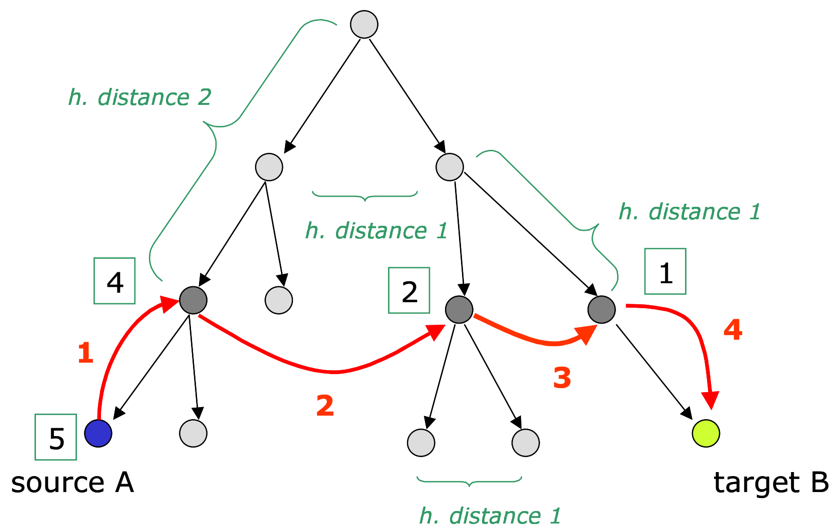

Figure 3: Example illustrating a search path using information about the target's position in the organizational hierarchy to direct a message. Numbers in the square give the h-distance from the target.

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 therefore had no e-mail records on our server), the mean number of steps was 4.7. This result indicates that not only are people typically easy to find, but nearly everybody can be found in a reasonable number of steps.

The last experiment we performed on the HP email network used the target's physical location. Figure 6 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 $(C - A) * 25' + (10 - 3) * 25' = 2 * 25' + 7 * 25' = 225$ 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 distance between two individuals' cubicles was naturally correlated to their separation in the organizational hierarchy, with the correlation coefficient $\rho = 0.35$. The fact that they are not aligned exactly shows that the physical locations do not strictly correspond to organizational layout. This can be due to several factors. For example, the cost of moving individuals from one cubicle to another when re-organizations occur or new individuals join, may outweigh the benefit of placing everyone in the same organizational unit in one location. The availability of other communication media, such as email, telephone, and instant messaging, reduces the frequency with which individuals need to interact face to face.

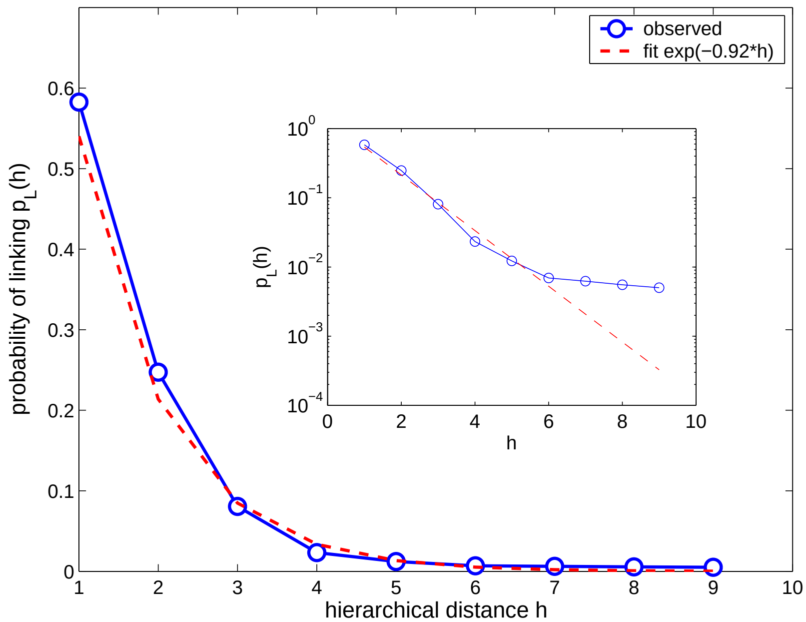

Figure 4: Probability of linking as a function of the separation in the organizational hierarchy. The exponential parameter $\alpha = 0.92$, is in the searchable range of the Watts model [13]

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 there is a tendency of individuals in close physical proximity to correspond. Still, individuals maintain disproportionately many far-flung contacts while not getting to know some of their close-by neighbors. The relationship between probability of acquaintance and cubicle distance r between two individuals, shown in Figure 7, 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 somewhere between $1/r$ and $1/r^2$. In any case, the observed $1/r$ relationship between distance and linking probability, implies a shortage of nearby links. This shortage hinders search 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.

Correspondingly, our simulations showed that geography could be used to find most individuals, but was slower, taking a median number of 7 steps, and a mean of 12. The fact that the strategy using geography trails behind a strategy using the target's professional position, is in agreement with Milgram's original findings. Travers and Milgram [11] divided completed chains between those that reached the target through his professional contacts and those that reached him through his hometown. On average those that relied on geography took 1.5 steps longer to reach the target,

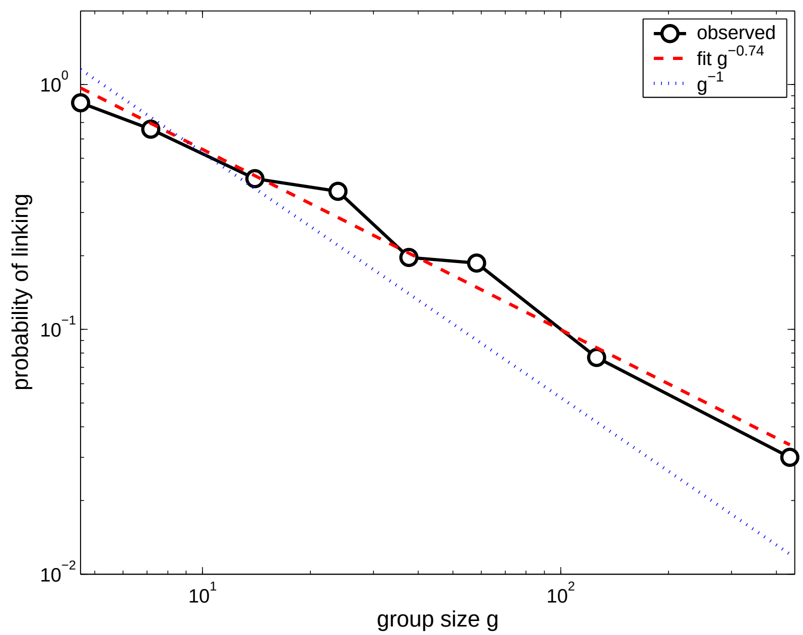

Figure 5: Probability of two individuals corresponding by email as a function of the size of the smallest organizational unit they both belong to. The optimum relationship derived in [7]is $p \sim g^{-1}$, g being the group size. The observed relationship is $p \sim g^{-3/4}$.

a difference found to be statistically significant. The interpretation by Travers and Milgram was the following: "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."

2.1 Summary of search results

Figure 8 shows a histogram of chain lengths summarizing the results of searches using each of the three strategies. It shows that both searches using information about the target outperform a search relying solely on the connectivity of one's contacts. It also shows the advantage, consistent with Milgram's original experiment, of using the target's professional position as opposed to their geographic location to pass a message through one's email contacts.

The simulated experiments on the e-mail network verify the models proposed in [13] and [6] to explain why individuals are able to successfully complete chains in 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.

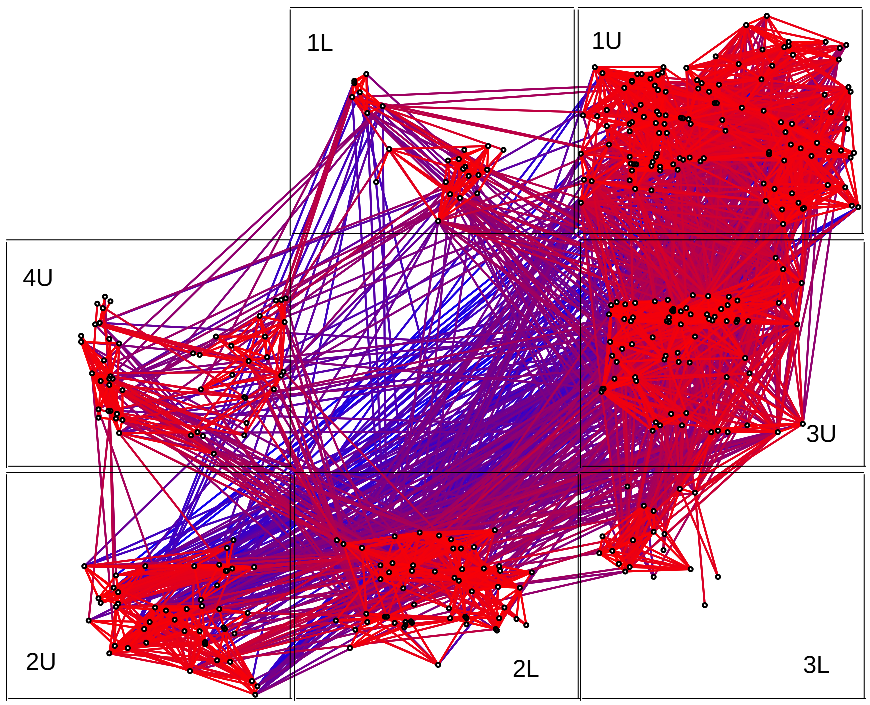

Figure 6: Email communications within HP Labs mapped onto approximate physical location based on the nearest post number and building given for each employee. Each box represents a different floor in a building. The lines are color coded based on the physical distance between the correspondents: red for nearby individuals, blue for far away contacts.

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 by Kleinberg in [6, 7], are not satisfied. The probability of linking is proportional to the inverse of the physical distance as opposed to the inverse of its square. The probability of linking is proportional to the group size to the negative three fourths, rather than being exactly inversely proportional. In both cases, there is an excess remote contacts and a shortfall of nearby contacts compared to what would be optimal.

Nevertheless, the probabilities of linking follow the correct overall tendency, and allow one to search this small community with a number of steps close to the shortest path. Hence the results of the email study are consistent with the model of Watts et al. [13]. 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 Figure 4, is well within the searchable regime identified in the model.

Our email study is a first step, validating these models on a small scale. It 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

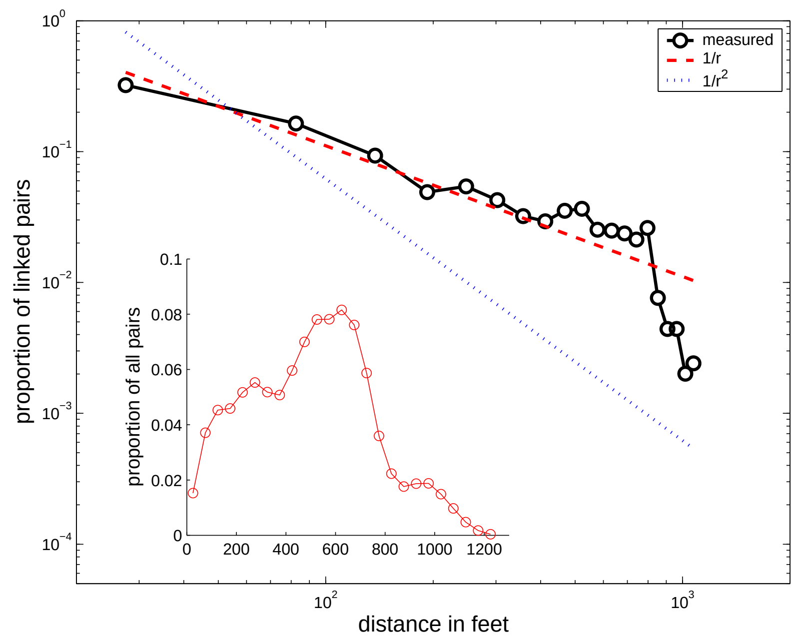

Figure 7: Probability of two individuals corresponding by email as a function of the distance between their cubicles. The inset shows how many people in total sit at a given distance from one another.

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.

3 Searching a network of Friends

E-mail data provides a fairly complete view of interpersonal communication within an organization and can be extracted automatically. In this section we are interested in exploring a different kind of network, one where each tie implies friendship as opposed to frequent communication. Email networks may not reveal individuals true affinities for one another, since some may reflect business communication between two people with no affinity for one another. At the same time, there is a chance that two people who consider each other friends do not email frequently. We obtained friendship network data from a community website, Club Nexus, that allowed Stanford students to explicitly list their friends.

The data gathered from the online community provides an opportunity to study how small world search works in this alternate social network. Over 2,000 undergraduate and graduate students joined Club Nexus and listed their friends as part of the registration process. The online community provided rich profiles for each of its users, including their year in school, major, residence, and gender. In addition, students selected words from a list to describe their personalities, how they like to spend their free time, and what they look for in friendship and romance. They could

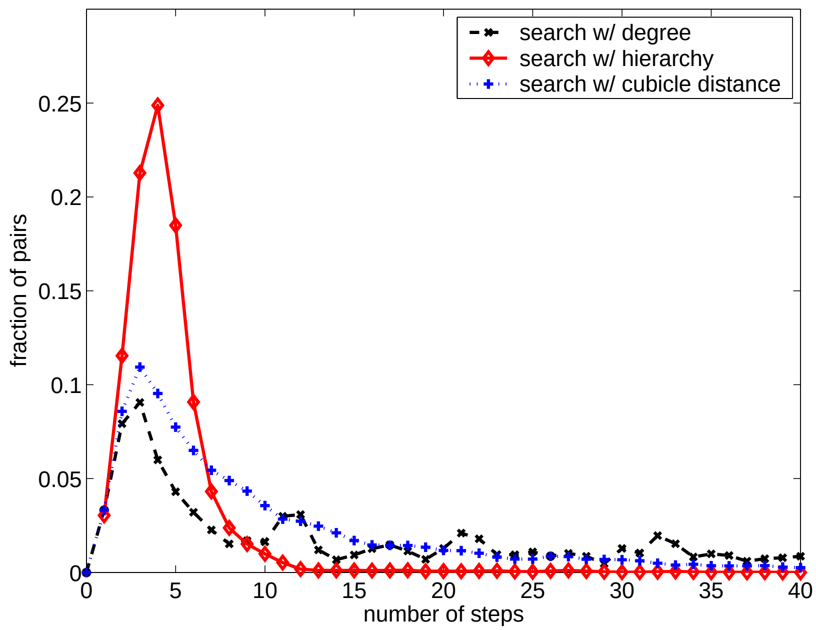

Figure 8: Results of search experiments utilizing either knowledge of the target's position in the organizational hierarchy or the physical location of their cubicle.

also optionally specify the book, movie, and music genres they liked, as well as sports and social activities that they enjoyed.

The richness of the profiles allowed for detailed social network analysis, including identifying activities and preferences influencing the formation of friendship (see [1]). However, the Club Nexus data differed in several respects from the HP Labs email network, and these differences made it difficult to apply a simple greedy strategy when searching the network.

The most important difference is the nature of the relationships revealed. In an email network, the relationship unambiguously reveals who talks to whom. While a website such as Club Nexus provides the opportunity to find links between individuals who do not explicitly communicate through a minable system (e.g. e-mail), it suffers from missing data. Users who are not considered close friends may not be linked to, and a link invitation may be ignored.

Approximately 209 users specified no friends, and a further 238 listed only 1 friend (probably the one friend who sent them an invitation to join the service). The full distribution is shown in Figure 9. Although many users listed only one or two friends, some had a 'buddy' list containing dozens of friends, resulting in an average degree of 8.2. This is in part a reflection of the fact that some users have more friends than others, but also that some are more eager to list their friends names, or list more than just their closest friends, on a website. To make the conditions for search more favorable, we recursively eliminated individuals with fewer than 3 friends from the network, leaving us with 1761 users, with median degree of 6, and average degree of 11.

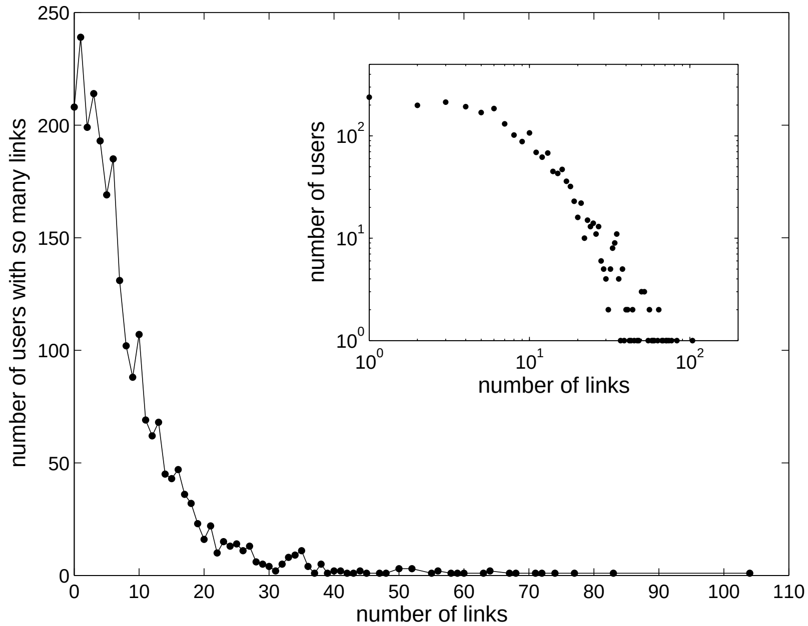

Figure 9: Distribution of the number of friends listed by Club Nexus users. The inset shows the same distribution on a logarithmic scale.

It is also important to point out that the Nexus Net reveals a particular kind of relationship, a friendship, and is hence more restrictive than an acquaintance network. It is possible that two female sophomore biology majors live in the same dorm, and interact at least occasionally because of the overlap in their location and interests, but do not link to each other on Club Nexus. We therefore do not have an accurate representation of who communicates with whom, but rather who considers whom a friend and is willing to say so in an online community.

A second handicap is the difficulty in creating a hierarchy with regard to most attributes. Physical location is given as a dorm, on-campus apartment complex, or off-campus apartment complex. The two on-campus graduate apartment complexes span a fourth to a half mile individually. Hence two graduate students living in the same apartment complex are far less likely to know one another than two undergraduates living in the same dorm and eating at the same cafeteria. Similarly, although some dorms cater to similar individuals (for example, all-freshman dorms or co-ops), and hence have higher incidence of people from one dorm being friends with people in the other, this is impossible to discern from the physical location of the dorms. Figure 10 shows the probability of two individuals being friends as a function of the distance between their residences (individuals living in the same residence have a distance of 0). Two people living in the same dorm have about a 5% probability of being Nexus 'buddies'. Two people living in different places have only a 0.3% chance of being buddies, regardless of how far apart these residences are. Hence a simple greedy geographical search, fairly successful in the case of the HP Labs email network, would not be able to hone in on a residence geographically on the Stanford campus.

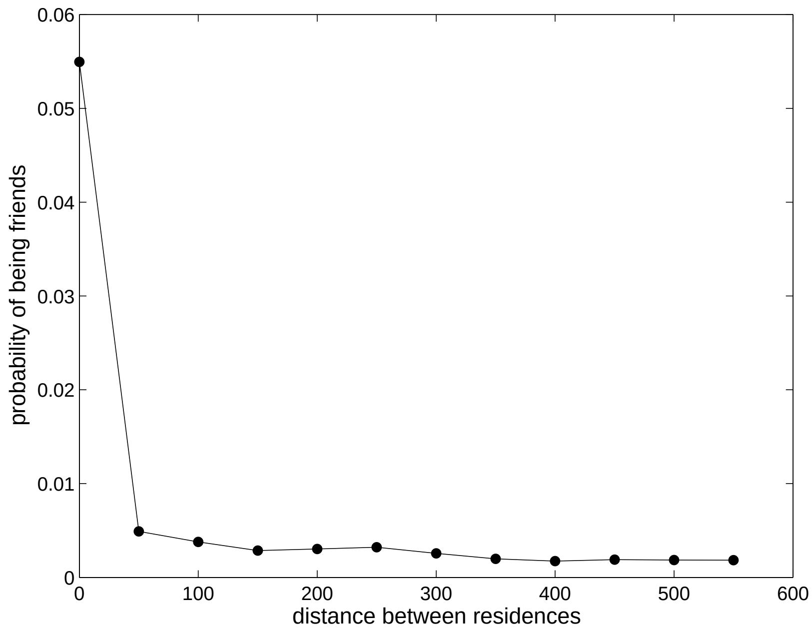

Figure 10: Probability of two individuals linking as a function of the separation between their residences

We encountered similar difficulties when attempting to form hierarchies for other attributes, such as department or year in school. We looked not only if users belong to the same department but also whether they belong to the same school. For example, two users might both belong to the School of Engineering, but one might be a Chemical Engineer while the other is a Mechanical Engineer. Likewise, if two individuals are not in the same year, we considered how many years separated them. Figure 11 shows how likely two individuals are to be registered as friends on Club Nexus, as a function of the number of years in school separating them. For both undergraduates and graduates two people in the same year have approximately a 1% chance of being Nexus 'buddies'. For individuals in different years that probability drops to less than 0.5%. This means that while people of similar age are more likely to know one another, that likelihood is still so small as to not be sufficient on its own to direct the search toward the target.

Among the many features collected by the Nexus website, only a few can be directly compared (or placed in a hierarchy). Thus, while physical location (distance between dorms) or year apart in school are comparable it is not clear how similar a swimmer is to a baseball or football player. Given the difficulty in placing most profile attributes, with the exception of year and department, into a hierarchy, we considered only whether an attribute is an exact match to the target's.

We found that using information about the target outperformed both a purely random strategy and one favoring well-connected nodes. A purely random strategy needed a median number of 133 steps, and and an average of 390. This strategy does not try to avoid nodes it has contacted before, and is clearly not well suited

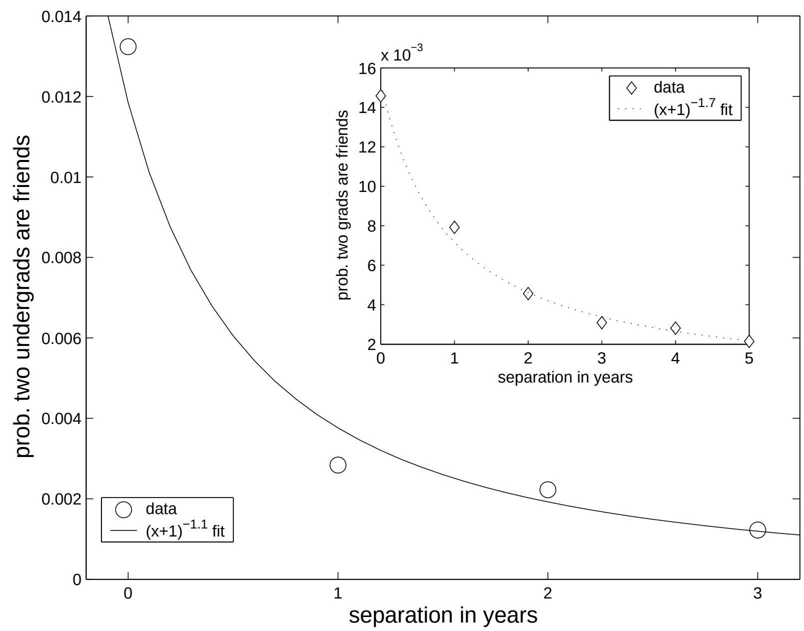

Figure 11: Probability of two undergraduates linking as a function of the difference in their year in school. The inset shows the same for graduate students.

to the network. A high degree strategy that discounts nodes already visited fared significantly better, taking a median number of 39 steps, and 137 on average. This number of steps is still too large to be practical, since the message is likely to be dropped before the long chain is completed.

The most successful search strategy compared 5 attributes simultaneously for the target and a person one is considering passing the message to. The possible combinations of attributes were:

- both undergraduate, both graduate, or one of each
- same or different year
- both male, both female, or one of each
- same or different residences
- same or different major/department

We then calculated the probability that two people know each other (or have an friend in common) based on all possible combinations of these variables. The simulation assumes that individuals would be able to judge, for example, the relative likelihood of someone in the same year knowing the target, as opposed to someone in a different year, but in the same dorm.

Including additional attributes, such as the number of shared interests in sports or other activities, did not significantly affect the search. Using the node profiles significantly outperformed the random and high degree strategies, requiring a median of 21 steps and a mean of 53. Taking twenty one steps is still a far cry from the six degrees found in real-world small world experiments, but it also comes close to becoming feasible. For example, if we follow the example in [13], assigning an attrition rate of $r = 0.25$, meaning that at every step there is a 25% chance that an individual will not pass the message on, then the network is just barely "searchable" with 5% of the messages reaching their target, in an average of 4.8 steps. However, one can imagine that if the network were larger, fewer than 5 percent of the messages would be able to reach their target due to the attrition rate and the number of steps required for searching.

What we see is that a simple greedy search is not particularly effective, especially once one is already confined to a small geographical space, such as a university campus. Further, the contacts that are available to the simulation are just a small fraction of individuals' actual contacts, and most of them are the individuals' closest friends. This is in contrast to the observation made by Granovetter [4], that it is the "weak ties" that play a disproportionately important role in bridging different portions of a social network. In the recent small world study, Dodds et al.[3], found relationships described as "casual" and "not close" to be more frequently used in successful chains.

In spite of the difficulties our algorithm encountered, there is little doubt that the small world experiment would succeed if conducted within the confines of Stanford University. First, students have many more than 11 friends on average. They are likely to know many of the individuals in the dorm, not just the 5% they listed with Club Nexus, as well as others in their own year attending the same classes. It is also likely that many individuals know their friends' friends, even if they choose not to associate with them directly. In the reverse small world experiment performed by Killworth and Bernard [5], participants chose on average about 210 different acquaintances to pass the message to various hypothetical targets. This indicates that individuals must be selecting from a large arsenal of acquaintances, something the Club Nexus data set does not provide.

The fact that participants in actual small world experiments have the opportunity to use information about their friends' friends and direct their search along different attributes of the target, is an important consideration for those constructing social software websites. Users only allowed to communicate with close, local neighbors will be at a disadvantage to those with access to information about their casual acquaintances and second degree neighbors. It is therefore important for the website to expose additional information or to suggest contacts based on a global view.

4 Conclusion

We simulated small world search in two scenarios. The first was within an organization, where a substantial portion of regular correspondence is captured through email. The second was an online community at a university where members of the community volunteer information about themselves and who their friends are. We constructed networks from both communities and simulated a simple greedy search on the network - each node passes the message to a neighbor who is most like the target. In the case of the email network, the strategies were successful - messages reached most individuals in a small number of steps, and using information about the target outperformed simply choosing the most connected neighbor. This was due in large part to the agreement with theoretical predictions by Watts et al. and Kleinberg about optimal linking probabilities relative to separation in physical space or in the organizational hierarchy.

In the case of the online community, strategies using information about the target were less successful, but still outperformed a simplistic high degree search. The limited success of greedy search is not surprising given that most available dimensions that could be searched on were binary, as opposed to organized into a hierarchy that would allow a search to hone in on its target. Geography was also almost a binary variable, since the probability of students being friends was on average independent on the separation between their dorms (unless they happened to live in the same one). Perhaps most importantly, non-participation or missing data biased the results by hiding connections that could be used in the search.

The social network structure of the real world probably lies somewhere in between - it is less structured than a small organization such as HP Labs, and at the same time people have many acquaintances based on their location, profession, age, and other interests. We can speculate that individuals participating in small world experiments are able to navigate geographical and professional space by using knowledge of a substantial portion of their immediate social networks and using more sophisticated strategies that look more than just one step ahead.

For the developers of social software it is important to understand how different data collection techniques (automated, implicit versus manual, explicit) impact the resulting social network and how these networks relate to the real world. Where the data is incomplete or reflects non-hierarchical structure, tools that support social search should assist users by either providing a broader view of their local community or directly assisting users through a global analysis of the network data.

Acknowledgements

We would like to thank Orkut Buyukkokten for the Club Nexus data and Joshua Tyler for help with obtaining cubicle distances at HP Labs. We would also like to thank Kevin Lai and Rajan Lukose for their helpful comments and suggestions.

References

[1] L. A. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the web. *First Monday*, 8(6), June 2003.

[2] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Phys. Rev. E*, 64:046135, 2001.

[3] P. S. Dodds, R. M., and D. J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 2003.

[4] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[5] P. Killworth and H. Bernard. Reverse small world experiment. *Social Networks*, 1:159–192, 1978.

[6] J. Kleinberg. Navigation in a small world. *Nature*, 406, 2000.

[7] J. Kleinberg. Small-world phenomena and the dynamics of information. *Advances in Neural Information Processing Systems (NIPS)*, 14, 2001.

[8] C. C. Lundberg. Patterns of acquaintanceship in society and complex organization: A comparative study of the small world problem. *Pacific Sociological Review*, 18:206–222, 1975.

[9] S. Milgram. The small-world problem. *Psychology Today*, 1:62–67, 1967.

[10] I. Pool and M. Kochen. Contacts and influence. *Social Networks*, 1:5–51, 1978.

[11] J. Traver and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[12] 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*. Kluwer Academic Publishers, Netherlands, 2003.

[13] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.

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

[15] F. Wu, B. A. Huberman, L. A. Adamic, and J.R. Tyler. Information flow in social groups. `http://arxiv.org/abs/cond-mat/0305305`, 2003.