

Friends and Neighbors on the Web

Lada A. Adamic and Eytan Adar

Xerox PARC
3333 Coyote Hill Rd.
Palo Alto, CA 94304
{ladamic | adar} @parc.xerox.com

Abstract

The Internet has become a rich and large repository of information about individuals. The links and text on a user's homepage to the mailing lists the user subscribes to are reflections of social interactions a user has in the real world. We devise techniques to mine this information in order to predict relationships between individuals. Further we show that some pieces of information are better indicators of social connections than others. The high quality information we discover provides a glimpse into the social life of two communities and has potential applications in automatically inferring real-world connections and discovering and labeling communities.

1. Introduction

One of the first large scale web applications was the serving of individual homepages. These generally autobiographical pages reflect a user's interests and experiences. They include anything from photographs of the user's pet to the user's essays or resume. Homepages are not free-floating in the web, but point to and are pointed at by other users. These links can represent anything from friendship, to collaboration, to general interest in the material on the other user's homepage. In this way individual homepages become part of a large community structure.

Recent work [6] [7] [9] has attempted to find of web pages communities by performing analysis on their graph structure. That is, given a graph, this method extracts clusters of users in the same "community." Rather than attempting to extract communities in our research we attempt to gain an understanding of the general structure of homepage based communities. By performing an analysis of *information side effects*, we describe and evaluate techniques that predict where connections between individuals will exist.

1.1 Information Side Effects

The harnessing of information side effects attempts to leverage information intended for one purpose in order to understand some tangential phenomena. A very nice example of

information side effects is the RadioCamera system [5]. RadioCamera mines information from cell phone base stations that show the load on any given tower in order to determine traffic conditions. Congested roadways will show a increased load on base stations than roads with no traffic.

Instead of traffic and cell phones we are interested in web structure and social networks. Authors of personal homepages intend to provide information about themselves and list a few of their friends. But there is an informational side effect of the collective action of many users doing this. Users linking to one another form a giant social network which is easy to harvest and provides a lot of information about the context of a link between two people.

Gathering information on relationships between people and the context of those relationships, which can range from cohabitation (i.e. fraternities) to shared interests (i.e. basketball), used is an arduous task for social networks researchers. Obtaining data requires time consuming phone or live interviews. We are able to harvest this information easily and automatically because it is already available as a side effect of people living a digital life. This presents an unprecedented opportunity to discover new and interesting social and cultural phenomena.

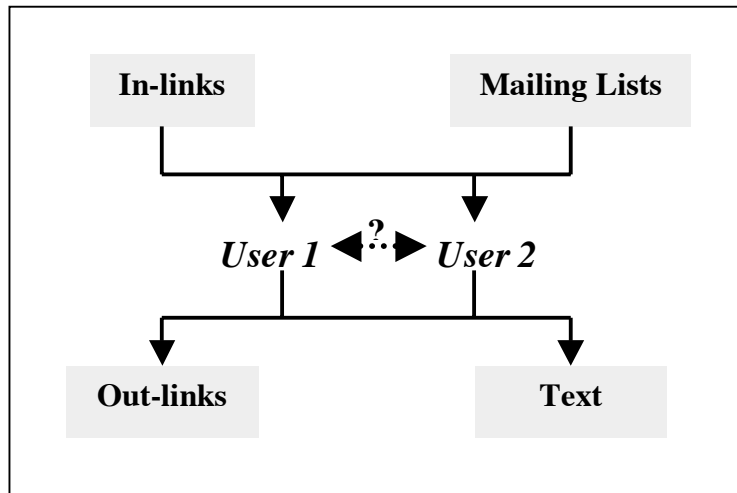


Figure 1 There were four sources of information for a user, in-links and mailing lists which were provided by external sources, and out-links and text which were provided by the users themselves. All four can be used as a means of inferring relationships between the users.

The data we study, as described below and in Figure 1, comes from the following four different sources:

- The *text* on user's home page provides semantic insight into the content of a user's page. Co-occurrence of text (we actually use multi-word "things" such as organization names, noun phrases, etc. instead of single word text) between users who link to each other usually indicates a common interest.
- *Out-links* are links from a user's homepage to other pages.

- *In-links* are links from other pages to the user's homepage. For example, a list of all members of a fraternity will link to individual homepages.
- *Mailing lists* provide us with valuable community structure that may not necessarily appear in homepage based communities.

In our case, we were interested in evaluating the ability of each of the above four sources of information to predict relationships between users. For example, we might expect that people associated with the same history class or the same fraternity might know each other. In order to uniformly evaluate these predictors it was necessary to build a constrained data set. We achieved this by crawling the home pages of students at Stanford University and the Massachusetts Institute of Technology (MIT), a process described in more detail below¹.

1.2 Paper Roadmap

In Section 2 of the paper we discuss community web page structures in terms of small world phenomena. Section 3 describes prediction schemes for link structures based on the information sources described above, and in Section 4 we discuss which particular types of information are useful for prediction in different domains. In Sections 5 and 6 we describe future work, potential applications of this technique, and draw general conclusions.

2. Link structure between personal homepages

It is well known that real world social networks are described by the small world phenomenon. This phenomenon is epitomized by the following situation: two people meet at a party and within minutes discover that they have an acquaintance in common. It appears to them that it's a small world, that is that everybody in the world is connected through only a short chain of acquaintances. The term 'six degrees of separation' first came up when Milgram [10] in the 1960's conducted an experiment asking a set of subjects in Nebraska to deliver a message to a specific target in Massachusetts. The subjects could pass the message only to someone they knew and who knew who they were as well. The message was passed along on average by only six people, a small number, considering that most people tend to move in close social circles tied to a geographic location, profession, or activity.

The structure of a small world network was mathematically formalized by Watts & Strogatz [11] who showed that social networks, such as the collaboration graph of film actors, are small world networks. It was subsequently shown that the world wide web [1] [3] is also a small world network. If social networks are small world graphs, and the web is a small world graph, then one would naturally expect networks of personal homepages

¹ All the information used in this analysis, with the exception of the MIT mailing lists, was publicly available. While we do not consider ourselves to be in violation of the spirit in which this information was made available, the potential for (ab)use of methods such as ours leads to an interesting set of ethical questions.

to be small world graphs as well. We confirmed this intuition by analyzing the networks of personal homepages at Stanford and at MIT and finding that they are small world networks as well.

It is quite popular for students to mention their friends on their homepages [12]. If those friends have homepages themselves, the student may include links to their friends' homepages. There are a number of social pressures for people to do this. As one of the authors discovered, once one person links to another, they expect a link in return. Those making a personal homepage for the first time may first look at their friends pages and notice that those friends have lists of friends. They imitate the practice of listing friends, and even pressure their friends into creating a personal homepage, just so that they can link to it.

We studied the phenomenon of listing friends on Stanford and MIT homepages. We looked at all users having a homepage under the domains `www.stanford.edu` and `{web,www}.mit.edu`. These are the homepages of students, faculty, and staff. Many students and faculty have personal homepages elsewhere, on departmental or personal machines or through external web-hosting. We omitted these pages, and crawled only pages under the specified domains looking for user to user links.

	Stanford	MIT
Users with non-empty WWW directories	7473	2302
Percent who link to at least one other person	14%	33%
Percent who are linked to by at least one other person	22%	58%
Percent with links in either direction	29%	69%
Percent with links in both directions	7%	22%

Table 1 Summary of links given and received among personal homepages at Stanford and at MIT

As Table 1 shows, about 30% of Stanford and 70% of MIT users with homepages are connected to other users, either by listing others or by being listed themselves. For this study, we chose to ignore the directionality of the links. That is, if one user links to another, we take it as evidence that the two people know each other. It is also safe to assume that the two people are friends, or at least have a professional relationship (for example, a student linking to their research advisor). There is a possibility that one user links to information on another's page without personally knowing the user. From our experiments we find that when this does happen it is easy to detect and those users are removed. For example, we found that many web pages were generated by modifying a template given out in introductory web design courses and contained links to the instructors' homepages. From here on we will use the term "friend" for any user who links to or is linked to by another.

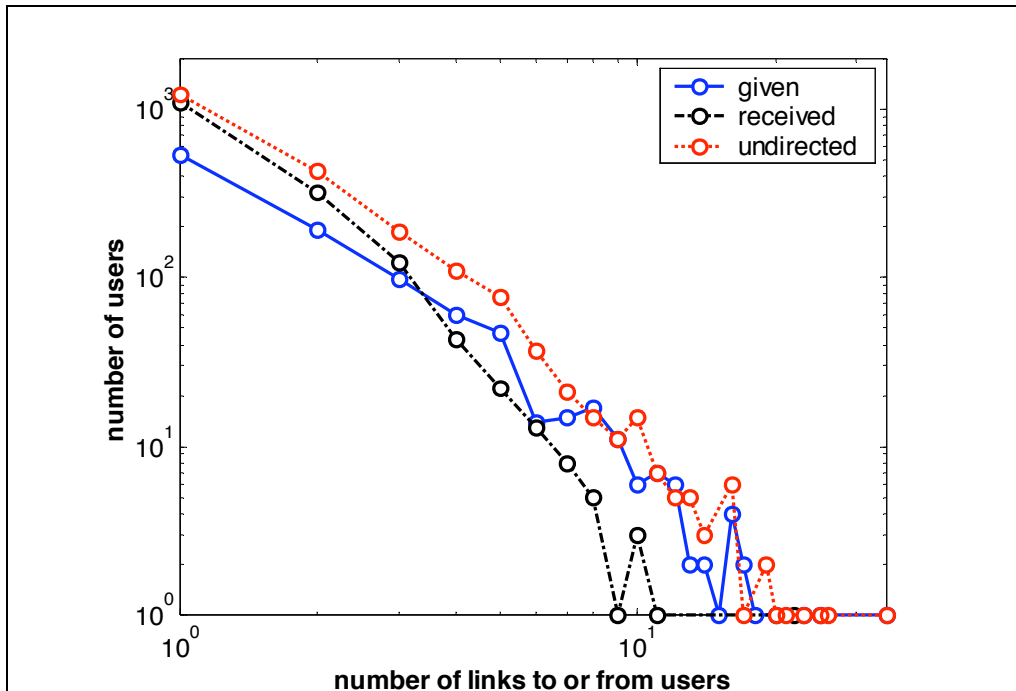


Figure 2 Distribution of given, received, and undirected links in the Stanford social web. Note the log-log scale. The averages were 2.5, 1.6, and 2.2 for given, received, and undirected links respectively.

Figure 2 shows the distribution of links either given or received between Stanford users on a log-log scale. The obvious pattern is one where users typically link to only one or two other users, with a very small but still significant fraction linking to dozens of users. This is also true of links *to* users. Some users are very popular, attracting lots links, while most get only one or two. The more startling result is that users linking to only 2.5 other people on average create a virtual connected social network of 1,265 people and a few smaller networks as summarized in Table. 2. We get a similarly sized network for MIT, even though the MIT set contained only a third of the number of homepages found at Stanford. This is due to a higher percentage of MIT users linking to one another as listed in Table 1.

Components size in users	Number of components		Percent of total users w/links	
	Stanford	MIT	Stanford	MIT
1281	0	1	0%	85.6%
1265	1	0	58.4%	0%
30-39	1	0	1.4%	0%
20-29	2	0	2.5%	0%
10-19	3	1	1.8%	0.8%
2-9	272	77	40%	13.6%

Table 2 Connected components of users linking to one another.

Figure 3a shows a layout of the graph of the largest set of connected users. There is a well-connected central core of users, surrounded by strands of less well connected users.

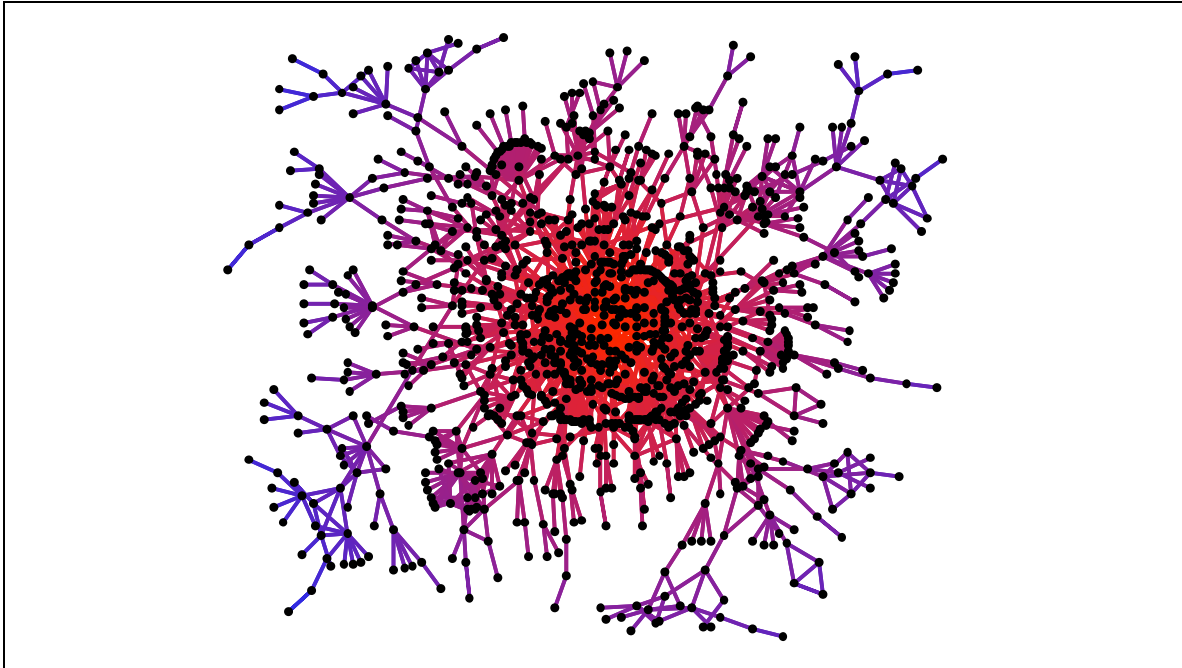


Figure 3a Graph layout of the Stanford social web.

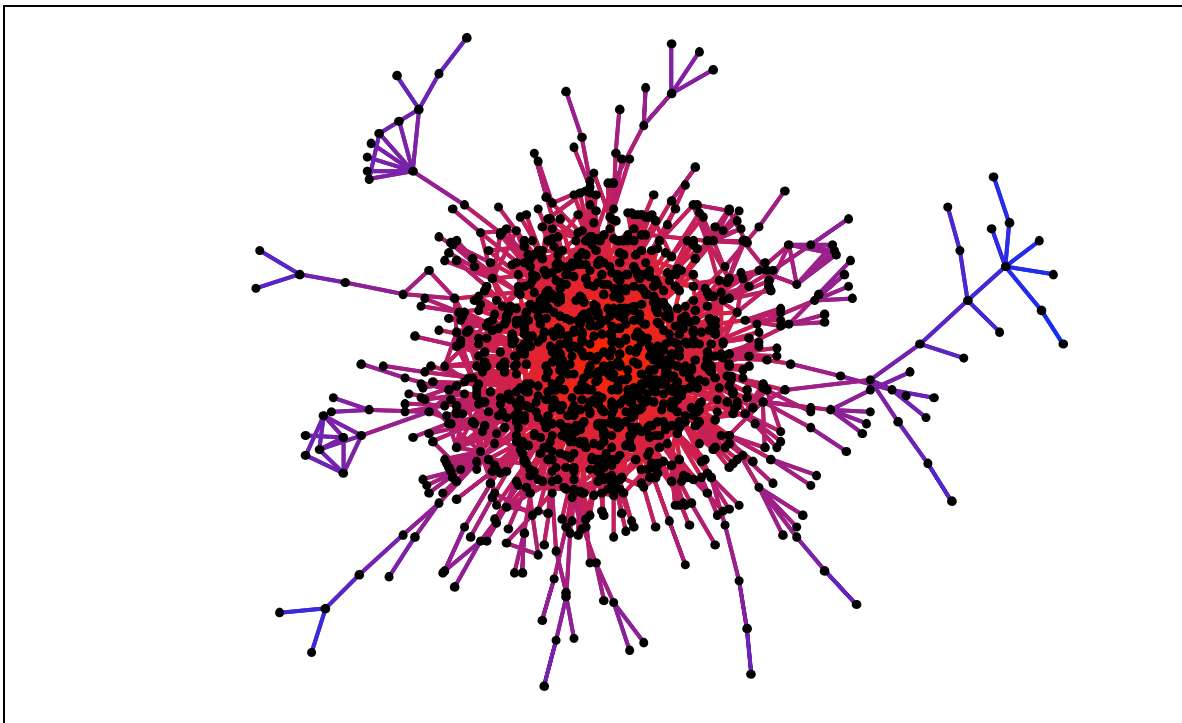


Figure 3b Graph layout of the MIT social web.

By the Watts and Strogatz definition of small worlds, the average shortest path from any two users in the large network is small, but there is still a lot of local structure (users within a group linking to one another). In the case of the Stanford social network, it is a mere 9.2 hops on average from one user to any other following links on users' pages. Comparing Figures 3a and 3b we see that MIT appears as a more tightly knit community. Indeed, this is reflected in the lower shortest average path of 6.4. These figures, already quite small, provide only an upper bound for the average shortest path between any two people at the university because people typically have many more friends and acquaintances than the ones they mention on their homepages.

The extent to which users band together can be measured via the clustering coefficient C . For a user who links to (or is linked to by) N other users, the clustering coefficient is the number of pairs of people out of the N who link to each other, divided by the number of all possible pairs ($N*(N-1)/2$). For the entire graph, C is obtained by averaging the individual coefficients for all the users. For the Stanford social web C turns out to be 0.22 while for MIT it is 0.21, both 70 times greater than for random graphs with the same number of nodes and edges. This means that if Jane links to Mary and Bob on her homepage, there is a 20% chance that either Mary links to Bob, or Bob links to Mary.

2.1 Context

The high degree of clustering among users allowed us to use 2D spectral embedding [4] techniques to visually cluster Stanford users. After identifying spatially distinct clusters, we viewed the homepages of the users to determine what those users have in common. Every cluster we looked at had some sort of common label: civil engineering professors, ITSS (information technology systems and services), or Greek electrical engineering grad students. The personal homepages revealed to us not only who whom, but why.

There were three different things we used to set a context for links:

1. Words and phrases the user mentions on their homepage
2. Links included in the user's homepage
3. External links to the user's homepage.
4. Email mailing lists the user belongs to.

We crawled each user's homepage and identified and extracted links to other users as well as external pages (those out-links pointing away from a user's pages). ThingFinder [8] was used to extract the words and phrases in the following categories: persons, places, cities, states, countries, organizations, companies, miscellaneous proper nouns, and noun groups. Because ThingFinder was designed with commercial applications in mind, it fares better in recognizing companies and organizations than things which might be more relevant to students such as hobbies and other interests. It is also fairly sensitive to capitalization, so that it might pick out "Social Networks", but not "social networks". Despite its shortcomings we used ThingFinder for its ability to categorize nouns and recognize phrases.

Complete lists of subscribers to mailing lists were obtained from a main mailing list server (mailing lists on departmental servers were not considered). Private lists could not be obtained. They comprised less than 5% of the total lists.

Finally, in-links were collected by querying Google (for Stanford) and AltaVista (for MIT) to obtain pages pointing at the individual's homepage. We required two different search engines due to the variety of URLs that all correspond to the same pages within MIT. AltaVista allowed for wildcard searches for links which Google did not.

We developed a web interface (available at <http://negotiation.parc.xerox.com/web10>) shown in Figure 4 that allows one to

- A. Find users with homepages by searching for names or browsing a directory
- B. Find text and links found on the user's homepage, as well as which mailing lists the user is subscribed.
- C. List whom the user links to and who links to them, then see what those users have in common
- D. Match a specific user to others based on links, text, and mailing lists.

Beyond developing the interface, we quantitatively evaluated the matchmaking algorithm for all three kinds of information about the user.

3. Predicting Friendship

We saw from the interface described above that people who link to each other usually have something in common. However, we wanted to reverse the situation and predict who is friends with whom by analyzing text, links, and mailing list. If we are trying to evaluate the likelihood that user A is linked to user B, we sum the number of items the two users have in common. Items which are unique to a few users are weighted more than commonly occurring items. The weighting scheme we use is the inverse log frequency of their occurrence. For example, if only two people mention an item, then the weight of that item is $1/\log(2)$ or 1.4, if 5 people mention the item, then its weight drops down to $1/\log(5)$ or 0.62. To summarize

$$similarity(A,B) = \sum_{shareditems} \frac{1}{\log[frequency(shareditem)]}$$

In order to evaluate the success of our friendship prediction scheme, we ranked the matches for each user in order of decreasing similarity separately for text, in and outgoing links, and mailing lists. Among the matches for each user, we identified the user's friends. Table 3 shows an example of our procedure. We measured the success of our procedure in terms of the placement of friends on the ranked list of matches. The measure is asymmetric with respect to a pair of friends. Person A can rank as 1st for person B, but person B might only rank 3rd for person A.

user 1: kpsounis Konstantinos Psounis	user 2:stoumpis Stavros Toumpis
Things in common	
CITIES:	Escondido, Cambridge, Athens
NOUN GROUPS:	birth date, undergraduate studies, student association
MISC:	general lyceum, NTUA, Ph.D., electrical engineering, computer science, TOEFL, computer
COUNTRIES:	Greece
Out links in common	
http://www.stanford.edu/group/hellas	Hellenic association
http://www.kathimerini.gr	Athens news
http://ee.stanford.edu	Electrical Engineering Department
http://www.ntua.gr	National Technical University of Athens
In links in common	
http://www.stanford.edu/~dkarali	Dora Karali's homepage
http://171.64.54.173/filarakia.html	Dimitrios Vamvatsikos friends list
Mailing lists in common	
greek-sports	Soccer/Basketball mailing lists for members of Hellas
hellenic	Hellenic association members
ee261-list	Fourier transform class list
ee376b	Information theory class list

Figure 4 Example of text, links, and mailing lists shared by two Stanford users linking to each other

If a friend had no item in common with the user, whether it be mailing list, text, or link, the friend would not be ranked. This can happen if we have very little information about one of the two users. It can also happen if the users use their homepages to express different interests. They might both share an interest in sports and beer, but one might devote his/her homepage entirely to beer, while the other devotes it only to sports. In this case we wouldn't be able to rank the friends with respect to each other based on out links or text because there would be no overlap. Table 4 shows what percentage of the time we were able to rank a friend.

anakken: Clifford Hsiang Chao		
Linked (friends)	Likeness Score	Person
NO	8.25	Eric Winston Liao
YES	3.96	John Andrew Vestal
NO	3.27	Desiree Dawn Ong
YES	2.82	Stanley Hsinheng Lin
NO	2.66	Daniel Sunil Chai
NO	2.55	Wei Nan Hsu
YES	2.42	David J. Lee
NO	2.41	Hans Christian Andersen
NO	2.41	Byung Joo Lee

Table 3 Top matches for a particular Stanford user, with the friends identified.

The number of terms, links, mailing lists, etc. a user has are distributed according to power-law [2], meaning that most people have only a few items, but a few have a large number, bringing the average higher overall. For example, for Stanford the average number of terms, out links, in links, and mailing lists were 113, 22, 3, and 6 respectively. In a power-law distribution, averages are usually much higher than the typical value or mode, but they nevertheless give a sense that people tend to include more text than links on their homepages. The number of terms recorded for a user was higher than the number of links. Hence we were able to make more matches with respect to terms, but the quality of those matches was not greater than that provided by the much less numerous links. In order to make a fair comparison between methods using each of the four types of information, we equalized the total number of matches made by introducing threshold value for which we would declare a match.

Figures 5a-d show how friends fared. They were more than twice as likely to be ranked 1st than 2nd, with the numbers decreasing from then on in a power-law fashion, as shown on the log-log plot in Figs. 4 b and d. This means that most frequently we predict the friends correctly, but every once in a while we give a friend a fairly low rank.

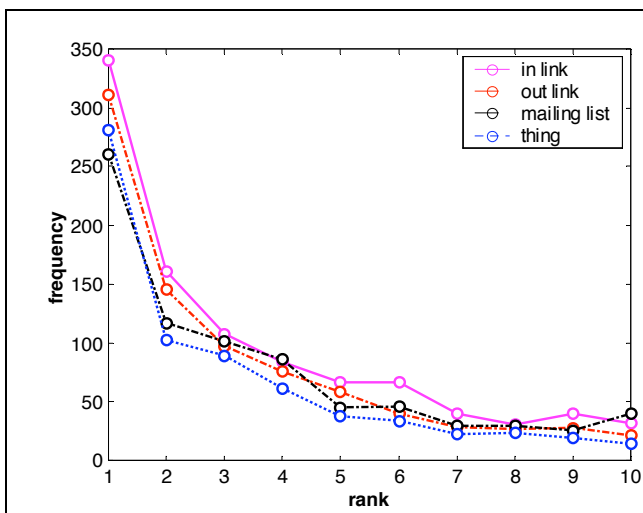


Figure 5a Linear scale plot showing how often we assigned each rank to a friend for the Stanford data

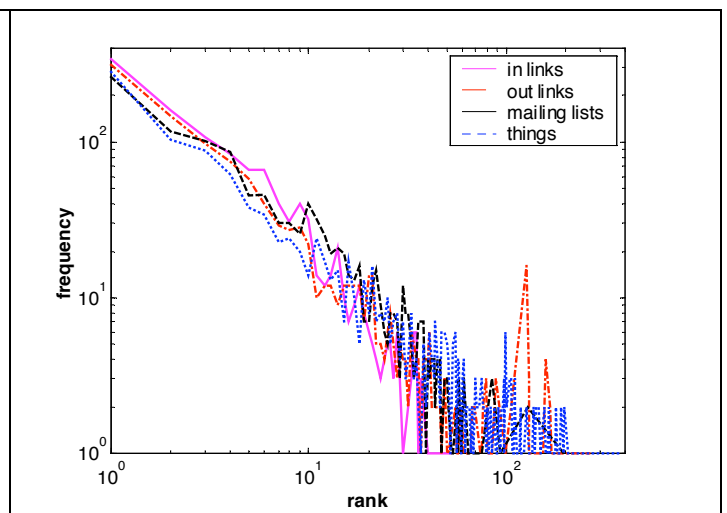


Figure 5b Stanford data on a log-log plot showing a power-law relationship.

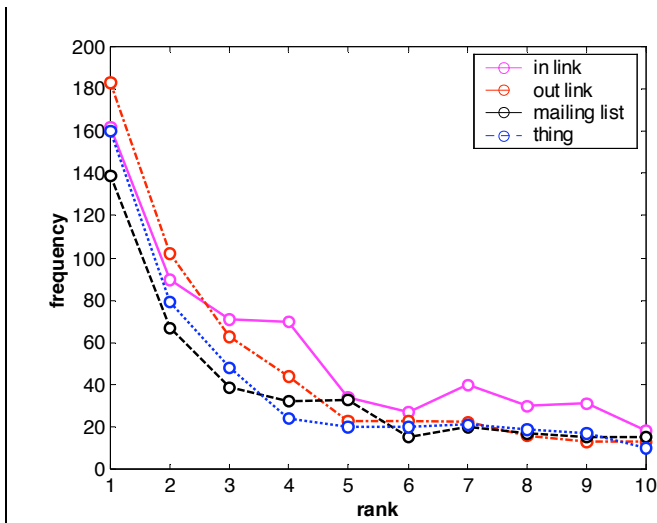


Figure 5c Linear scale plot showing how often we assigned each rank to a friend for MIT data

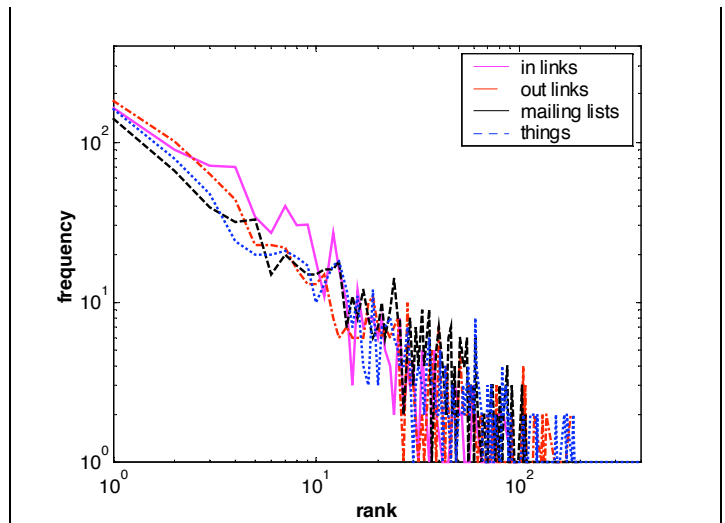


Figure 5d MIT data on a log-log plot showing a power-law relationship.

Table 4 gives a summary of the results. Links and mailing lists prove to identify friends more reliably than text.

Method	Pairs ranked Stanford	Pairs ranked MIT	Average rank Stanford	Average rank MIT
in-links	24%	17%	6.0	9.3
out-links	35%	53%	14.2	18.0
mailing lists	53%	41%	11.1	22.0
text	53%	64%	23.6	31.6

Table 4. Coverage and the ability to predict user-to-user links for 4 types of information about the user. The average rank was computed for matches above a threshold such that all 4 methods ranked an equal number of users.

4. Individual links, terms, and mailing lists as predictors

So far we have used a simple weighting scheme for items associated with a user which measured only how commonplace the item was. Intuitively one would expect some items to be shared only by friends, while others could be associated with almost anyone. We measured an individual item's ability to predict whether two people who mention it will link to one another. The metric we used was simply the ratio of the number of linked pairs of users who are associated with the item, divided by the total possible number of pairs, given by $N/(N-1)$, where N is the number of users associated with the item.

4.1 Text

Some things are obviously more frequently shared by friends than others. Academic departments, research groups, dorms, religion, sports, and activities are more likely to be shared by friends than strangers. On the other hand, some things like major cities, software applications or common words are shared by disparate sets of people.

Within both MIT and Stanford the best predictors for linking are specific user names. Excluding terms labeled as names we find a very consistent set of terms that have high predictive power. For MIT we list the top 10 terms which are included in the table below. This table illustrates the tremendous predictive power of living group names (6 of the top 10). This feature, which is unique to the MIT data, is most likely the result of the tight integration of fraternities into the MIT social structure and is absent from the Stanford data.

Term (description)	Pairs	Users	Ratios
Union Chicana (student group)	5	5	.5
Phi Beta Epsilon (fraternity)	6	7	.286
Bhangra (traditional dance, practiced within a club at MIT)	4	6	.267
neurosci (appears to be the journal Neuroscience)	5	7	.238
Phi Sigma Kappa (fraternity)	5	7	.238
PBE (fraternity)	8	9	.222
Chi Phi (fraternity)	7	11	.127
Alpha Chi Omega (sorority)	4	11	.727
Stuyvesant High School	5	13	.064
Russian House (living group)	5	14	.055

Table 5 Top 10 term predictors mentioned by 5 or more users.

Table A1 (in the Appendix) lists the top 10 predictors among the terms found on personal homepages omitting proper names for Stanford. Here we find a group of engineering students from the same Greek university who link to each other, followed by other student groups: an Asian mentoring group, a marching band section, an undergraduate Asian American association, and only one sorority.

In contrast to the above terms which are linked to specific small communities, there are general terms which are mentioned by dozens of people, but those people don't link to one another. Some organizations, software, cities, and degree titles fall among these: NASA, congress, NT, BA, M.A., Atlanta, and Oakland.

4.2 Out-links

Out-links were analyzed in a similar fashion to terms. Users' web pages were crawled, links extracted, and then tested for their significance with respect to pairs of people linking to one another.

We find that the best predictors for linking appear to be links referring to another individual's home page. This friend of a friend phenomenon, corresponding to the high clustering coefficients of both social networks, is quite striking. Of the top 20 predicting links in the MIT data we find that 19 are links to individual homepages (10 out of the first 10). For Stanford, 9 out of the top ten links were also individual homepages. Table A2 lists other types of out-links among the top predictors for Stanford and shows that small

cohesive groups such as sororities, fraternities, research labs, and ethnic groups are strong predictors for friendship.

Removing these links, we list the 10 best out-link predictors in Table 6:

Page	Pairs	Users	Ratios
MIT Campus Crusade for Christ*	3	4	.5
The Church of Latter Day Saints	4	5	.4
The Review of Particle Physics	2	4	.333
New House 4 (dorm floor, home page)*	2	4	.333
MIT Pagan Student Group*	2	4	.333
Web Communication Services*	2	4	.333
Tzalmir (role playing game)*	2	4	.333
Russian house (living group) comedy team *	2	4	.333
Sigma Chi (fraternity)*	2	4	.333
La Unión Chicana por Aztlán	2	4	.333
*Pages within MIT			

Table 6 MIT link predictors mentioned by 4 or more users.

Predictably, links that perform poorly in identifying potential user connections consist mostly of popular general interest sites that are not specific to any particular group of people. The worst ranked sites include search engines and portals (Altavista, Lycos, Yahoo, etc.), as well as general MIT and Stanford pages (top level homepages such as www.mit.edu and www.stanford.edu).

4.3 In-Links

Just as in the out-link scenario above we find that a very significant portion of the top predictors for are individual home pages (5 out of 10 for MIT and 9 out of 10 for Stanford). This is consistent with the idea that if user A is friends with B and C there is a large chance that B and C will be friends as well. Other pages that appear to predict links are living groups and tightly knit religious groups with a web presence. Table 7 below is the top 10 in-link predictors for MIT (Table A3 is the equivalent table for Stanford).

Page	Pairs	Users	Ratios
Individual user's homepage links	5	4	.833
Freshman members of MIT Fraternity*	4	4	.667
Individual user's homepage links*	6	5	.6
Contributors to MIT web based art project*	3	4	.5
Residents of one unit (floor) of a dorm*	3	4	.5
Students of a small class*	3	4	.5
Individual user's homepage links*	3	4	.5
Members of religious group*	3	4	.5
Individual user's homepage links	3	4	.5
Individual user's homepage links*	3	4	.5

Table 7 Top 10 in-link predictors for MIT.

Lists with a large number of unrelated participants made poor predictors. They included general home pages, alumni lists reflecting many graduating years, and lists of all students in a major.

4.4 Mailing Lists

In the case of MIT, living group and religious organizations dominate the top predictive mailing lists. Table 8 represents the top 10 predictive mailing lists. We have opted to remove the specific list names and information as this information that can only be gathered from access to MIT servers from this table.

List description	Pairs	Users	Ratios
Summer social events for residents of specific dorm floor	9	8	.3214
Religious group	9	10	.2
Religious group	9	11	.1636
Religious group	4	8	.1429
Intramural sports team from a specific dorm	4	8	.1429
Summer social events for residents of specific dorm floor	19	17	.1398
Religious a cappella group	5	9	.1388
Intramural sports team from a specific dorm	8	12	.1212
"...discussion of MIT life and administration."	5	10	.1111
Religious group	10	14	.1099

Table 8 MIT mailing list predictors with 8 or more subscribers.

In Stanford (represented in Table A4) the best predictors for friendship consist mostly of lists specific to dorms, research groups, and ethnic, cultural and other small cohesive groups.

For MIT, the mailing lists that appear to be bad predictors fall into 2 main categories: very general discussion lists, and social activities. The social activities category is by far the largest, and appears to consist primarily of non-housing unit based lists. This category contains, for example, general announcements for a theater group and other sports lists such as scuba diving and cycling. In the discussion category we find lists related to programming questions and mailing lists dedicated to general discussions in large living groups.

The poor predictors for Stanford were similar to MIT. They included overseas program lists (*gosantiago*, *goflorence*), individual sports (*stanford_cyclists*, *snowboarders*, *runningclub*) and large department mailing lists: *suse-jobs* (school of education jobs list), *gsb-unofficial* (graduate school of business), *sms-news* (school of medicine news).

5. Future Work

We interact with dozens of people on a regular basis, but do not always put links to all of them. Users presented with their best matches given by our algorithm frequently recognize the individuals listed, even if they had not expressly put a link to them from their homepage. The fact that we do not have a complete list of friends results in many false negatives. That is, we correctly match a user to someone they know but we have no explicit link confirming this relationship. This makes a complete evaluation difficult, as measures such as precision-recall rely on a complete data set (list of friends in our case).

We would like to conduct a user study to obtain this information, for example whether the match is a close friend, acquaintance, or a stranger to the user. The results would extend our set of known user relationships and allow us to more accurately evaluate our matching algorithms.

6. Conclusions

We have shown that personal homepages provide a glimpse into the social structure of university communities. Not only do they reveal to us who knows whom, but they give us a context, whether it be a shared dorm, hobby, or research lab. Obtaining data on social networks used to be a tedious process of conducting a series of phone or live interviews. Studying social networks online can give us rich insight into how social bonds are created, but requires no more effort than running a crawler on home pages.

In this study we have demonstrated a means of leveraging text, mailing list, in and out-link information to predicting link structure. From least to most predictive these are: text, mailing lists, out-links, and in-links². We have also characterized specific types of items from each of these categories which turn out to be good or bad predictors.

Among the numerous applications of these results is the mining of correlations between groups of people, which can be done simply by looking at co-occurrence in homepages of terms associated with each group. Using these techniques in combination with community discovery algorithms yields labeled clusters of users. Thus, not only is it possible to find communities, but we can describe them in a non-obvious way.

Another possible application is the facilitation of networking within a community. Knowing which of friend of a friend is involved in a particular activity can help users find a chain of acquaintances to reach the people they need to. Finally, networks of homepages open a whole range of possibilities in marketing research, from identifying which groups might be interested in a product to relying on the social network to propagate information about that product.

² The rank of mailing lists and out-links is reversed between MIT and Stanford.

Acknowledgements

The authors would like to thank Bernardo Huberman and Rajan Lukose for their advice and comments.

References

- [1] L. Adamic, "The small world Web," *Proceedings of the European Conf. on Digital Libraries*, 1999.
- [2] L. Adamic and Eytan Adar, "Frequency of friendship predictors," <http://www.parc.xerox.com/iea/papers/web10/>
- [3] R. Albert, H. Jeong, A.-L. Barabasi, "The diameter of the World Wide Web," *Nature* 401, 130 (1999).
- [4] G.S. Davidson, et al, "Knowledge Mining with VxInsight Discovery Through Interaction," *Journal of Intelligent Information Systems*, Volume 11, Number 3, November/December, 1998, pp.259-285.
- [5] S. Diaz, "Cell Phone Signals Touted to Fight Traffic Wars," *San Jose Mercury News*, Jan. 20, 2000, <http://www0.mercurycenter.com/svtech/news/indepth/docs/traf012100.htm>.
- [6] G. Flake, S. Lawrence, and C. Lee Giles. "Efficient identification of web communities". In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, August 20-23 2000, pp.150-160.
- [7] D. Gibson, J. Kleinberg, and P. Raghavan. "Inferring Web communities from link topology," *Proceedings 9th ACM Conference on Hypertext and Hypermedia*, 1998
- [8] InXight ThingFinder product page, http://www.inxight.com/products_wb/tf_server/index.html.
- [9] R. R. Larson, "Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace," *Global Complexity: Information, Chaos and Control*, the 1996 Annual Meeting of the American Society for Information Science, October 21-26, 1996, Baltimore, Maryland, USA.
- [10] S. Milgram, "The small world problem," *Psychology Today* 1, 61 (1967).
- [11] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature* 393, 440 (1998).
- [12] Patricia Wallace, *The Psychology of the Internet*, Cambridge University Press, Cambridge, 1999

Appendix A: Top 10 predictors for Stanford

Term (description)	pairs	users	ratio
NTUA (National Technical University of Athens)	4	4	0.667
Project Aiyme (mentoring Asian American 8 th graders)	4	4	0.667
pearl tea (popular drink among members of a sorority)	7	6	0.466
clarpic (section of marching band)	4	5	0.4
KDPhi (Sorority)	6	6	0.4
technology systems (computer networking services)	4	6	0.267
UCAA (Undergraduate Asian American Association)	4	6	0.267
infectious diseases (research interest)	5	7	0.238
viruses (research interest)	10	12	0.152
home church (Religious phrase)	5	9	0.139

Table A1 The top 10 terms shared by the greatest number of people who mention them. The table omits people's names, which were by far the most successful predictors of a relationship between two people (people who mention the same third person are also more likely to link to each other). Terms mentioned by fewer than 4 people were omitted.

Page	Pairs	Users	Ratio
alpha Kappa Delta Phi (Sorority)*	19	7	0.9
National Technical Univeristy Athens	9	7	0.43
Ackerly Lab (biology)*	4	5	0.4
Hellenic Association*	8	7	0.38
Iranian Cultural Association*	4	6	0.27
Mendicants (a cappella group)*	7	8	0.25
Phi_Kappa_Psi (fraternity)*	6	6	0.21
Magnetic Resonance Systems Research Lab*	9	10	0.2
Applications assistance group*	9	11	0.16
ITSS instructional programs*	25	10	0.55
* Pages within Stanford			

Table A2 The top 10 Stanford predictor out-links. Only links mentioned by 4 or more people were considered. The list excludes personal homepages which otherwise dominate the top 10.

List	Pairs	Users	Ratio
Individual's list of friends	7	5	.7
Individual's list of friends	4	4	.667
Individual's list of friends	4	4	.667
Individual's list of friends	4	4	.667
Individual's list of friends	4	4	.667
Individual's list of friends	4	4	.667
Individual's list of friends	4	4	.667
Individual's list of friends	4	4	.667
Individual's list of friends	4	4	.667
Sorority member list	28	10	.622

Table A3 Top 10 in-link predictors for Stanford.

List	Pairs	Users	Ratio	Description
kairos97	8	9	0.111	dorm
mendicant-members	4	9	0.055	a cappella group
cedro96	5	11	0.045	dorm summer mailing list
first-years	9	15	0.042	first year economics doctoral students
local-mendicant-alumni	3	9	0.042	local a cappella group alumni
john-15v13	3	9	0.042	Fellowship of Christ class of 1999
stanford-hungarians	7	14	0.038	Hungarian students
serra95-96	2	8	0.035	dorm
metricom-users	3	10	0.033	network services employees who use metricom
science-bus	9	17	0.033	science education program organized by engineering students

Table A4 Stanford lists with 8 or more subscribers which serve as best predictors for users linking to one another.