Friends and Neighbors on the Web

Lada A. Adamic
Xerox PARC
3333 Coyote Hill Rd.
Palo Alto, CA 94304
(650) 812-4778
ladamic@parc.xerox.com

Eytan Adar
HP Sand Hill Laboratory*
1501 Page Mill Road M/S 1U-19
Palo Alto, CA 94304
(650) 857-2398
eytan@hpl.hp.com

ABSTRACT
The Internet has become a rich and large repository of information about us as individuals. Anything from 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. In this paper 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, and that these indicators vary between user populations and provide a glimpse into the social lives of individuals in different communities. Our techniques provide potential applications in automatically inferring real-world connections and discovering, labeling, and characterizing communities.

Keywords
Homepage analysis, social network, small worlds, web 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, our "friends and neighbors" on the web. 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] [10] has attempted to use analysis of link topology to find "web communities." These web communities are web page collections with a shared topic. For example, any page dealing with 'data mining' and linking to other pages on the same topic would be part of the data mining page collection. Such a page is not necessarily a homepage or even associated with a particular individual. In contrast, our work focuses on *individuals'* homepages and the connections between them, essentially allowing us to tap into both virtual and real world communities of people.

Although homepage identification has been researched as a separate problem [8][12], to our knowledge this is the first link analysis on a network of homepages. Rather than discarding the previous concept that pages which share a topic are likely to link to one another, we can now use it to characterize relationships between people. For example, are people who mention 'dance troupe' likely to link to each other? Consequently, can we use terms on homepages to predict where connections between individuals will exist? And furthermore, which terms are best at predicting connections: is 'dance troupe' a better predictor than 'kayaking'? Here we describe and evaluate techniques to answer the above questions. While the intent of homepages is to provide a view of the individual user and their local relationships to others, as a side effect they provide an interesting view of whole communities[1].

1.1 Information Side Effects
Information side effects are by-products of data intended for one use which can be mined in order to understand some tangential, and possibly larger scale, phenomena. A nice example of information side effects is the RadioCamera system [4]. 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.

Just as it is possible to extract global traffic patterns from a device intended to provide communication between two individuals, we can likewise extract large social networks from individualized homepages. 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 individuals.

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), is an arduous task for social networks researchers. Data is acquired through 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.

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

[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.

*Work done while author was at Xerox PARC

1. **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.

2. **Out-links** are links from a user's homepage to other pages.

3. **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.

4. **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 below.

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 communities. In Sections 5 and 6 we provide areas for future work, potential applications of this technique, and draw general conclusions.

2. HOMEPAGE LINK STRUCTURE

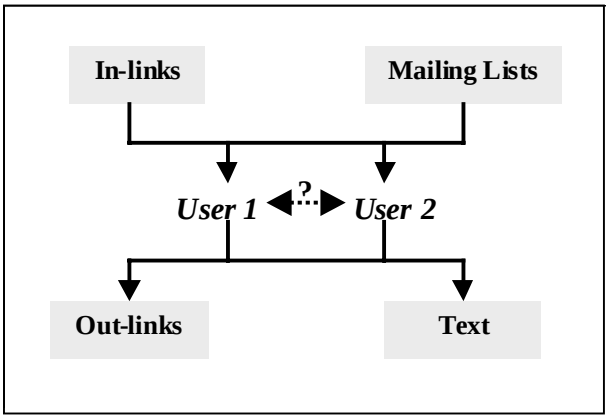

Figure 1 There are four sources of information for a user: in-links and mailing lists which were provided by external sources, and out-links and text which were provided by the users themselves. All four can be used as a means of inferring relationships between the users.

Real world social networks are described by the small world phenomenon. This phenomenon is familiar to anyone who has said 'It's a small world, isn't it?', upon discovering a mutual acquaintance shared with a stranger. It appears to them that everybody in the world must be connected through only a short chain of acquaintances. Social psychologist Stanley Milgram [11] in the 1960's tested the phenomenon experimentally by asking a set of subjects in Omaha, Nebraska to deliver a message to a specific target in Boston, Massachusetts. The participants could pass the message only to people they knew on a first name basis, and yet the message was passed an average of only six times. This coined the term 'six degrees of separation', 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 and Strogatz [13] to be a graph with a small average shortest path, and high cliquishness. They also 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 is also a small world network [1][3]. Given that both social networks and the web are small world graphs, we expected networks of personal homepages to be small world graphs as well. We confirmed this intuition by analyzing the networks of personal homepages at Stanford and at MIT.

Homepage networks arise because it is popular for students to mention their friends on their homepages [12], and link to those friends' homepages if they exist. They might be imitating lists they've seen on their friend's homepages, or they might even have been talked into creating a homepage, just so that their friends could link to it.

For this study, we looked at all users having a homepage under the domains www.stanford.edu and {web,www}.mit.edu. These

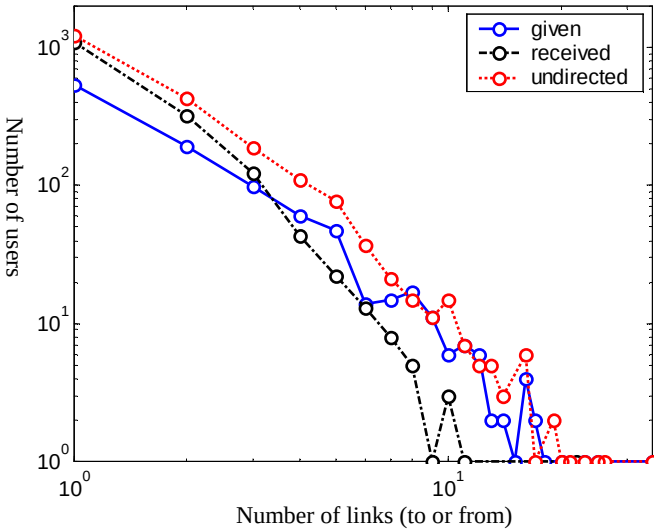

Figure 2 Distribution of given, received, and undirected links in the Stanford social web. Note the log-log scale. The averages were 2.5, 1.6, and 2.2 for given, received, and undirected links respectively.

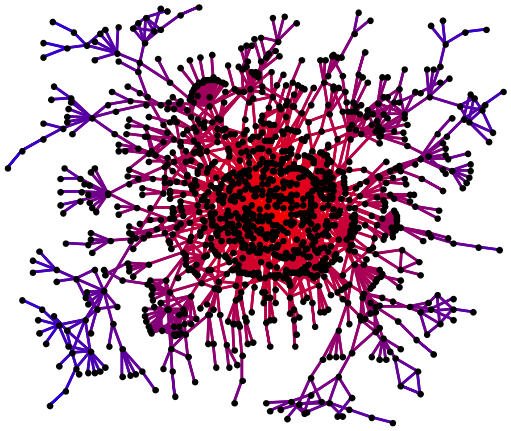

Figure 3a: Graph layout of the Stanford social web. Each node is an individual and each edge is a connection corresponding to a link between the two individual's homepages.

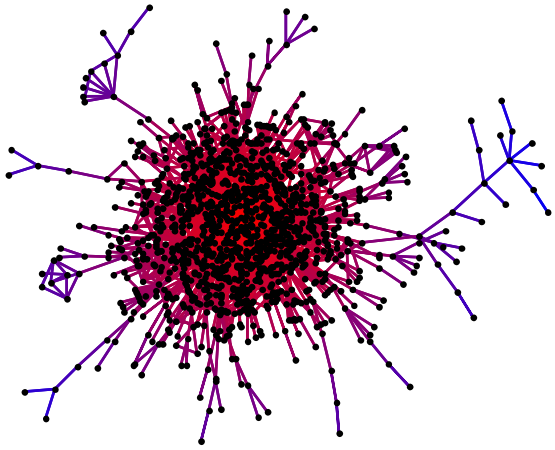

Figure 3b: Graph layout of the MIT social web.

sites contain 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. For simplicity, we omitted these external pages, and crawled only pages under the specified domains looking for user to user links.

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

	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%

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 at Stanford were generated by modifying a template given out in introductory web design courses and contained links to the instructors' homepages. These links were removed from the data set. From here on we will use the term "friend" for any user who links to or is linked to by another.

Figure 2 shows the distribution of links either given or received between Stanford users on a log-log scale. Users typically provide out-links 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 in-links to users. Some users are very popular, attracting many links, while most get only one or two. The link distributions correspond to real world social networks, where some people maintain a large number of active contacts or are very popular, but most people maintain just a select few friendships. 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 accounting for 58% of the users and a few smaller networks making up the remainder. At MIT, a full 85.6% (1281 users) belong to the giant component. This is due to a higher percentage of MIT users linking to one another as listed in Table 1.

Figures 3a and b show a layout of the graph of the largest sets of connected users for Stanford and MIT. There is a well-connected central core of users surrounded by strands of less well connected users. In the case of the Stanford social network, the average shortest path is a mere 9.2 hops 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[2].

[2] It is important to realize that web links only reflect a subset of the actual social network. While the number of hops may seem larger than previous experiments they only reflect an upper bound on this statistic.

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. These high clustering coefficients, combined with the small average shortest paths, identify both the MIT and Stanford social networks as small world networks.

2.1 Context

While link structure provides an interesting view of the social network in homepage communities it does not necessarily provide us with an understanding of why these links exist and how we may predict them.

To automate the task of giving links context we gathered four types of data: text, out-links, in-links and mailing lists. Text and out-links (including links to other users) were extracted from crawls of each user's homepage. ThingFinder [8] was used to extract the words and phrases in the text in the following categories: persons, places, cities, states, countries, organizations, companies, miscellaneous proper nouns, and noun groups. While ThingFinder is an improvement over using single terms it was designed with commercial applications in mind. Thus, it fares better in recognizing companies and organizations than phrases and names which might be more relevant to students such as hobbies or majors. It is also fairly sensitive to capitalization, so that it might pick out "Social Networks", but not "social networks". Despite its minor shortcomings, ThingFinder worked well for the homepage data we obtained.

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 at Stanford.

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) that allows users to:

A. Find individuals with homepages by searching for names or browsing a directory

B. Find text and links found on a user's homepage, as well as which mailing lists the user is subscribed to.

C. List whom the user links to and who links to them, then see what those users have in common (as illustrated in Figure 4)

user 1: **kpsounis**	**user 2**:**stoumpis**
Konstantinos Psounis	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 output of the person-to-person likeness program. The various terms, links, and mailing lists that two users have in common are shown.

D. Match a specific user to others based on links, text, and mailing lists. The algorithm for which is described below.

3. PREDICTING FRIENDSHIP

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

To predict whether one person is a friend of another, we rank all users by their similarity to that person. Intuitively, our matchmaking algorithm guesses that the more similar a person is, the more likely they are to be a friend.

Similarity is measured 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 that 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:

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

It is possible with this algorithm to evaluate each shared item type independently (i.e. links, mailing lists, text) or to combine them together into a single likeness score.

3.1 Evaluation

We evaluate the performance of the algorithm by computing the similarity score for each individual to all others, and rank the others according to their similarity score. We expect friends to be more similar to each other than others, and we measure this in two steps. First, we see how many of the friends can be ranked at all. That is, we compute what fraction of friends have a non-zero similarity score. Second, we see what similarity rank friends were assigned to.

Friends can appear to have no items in common 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.

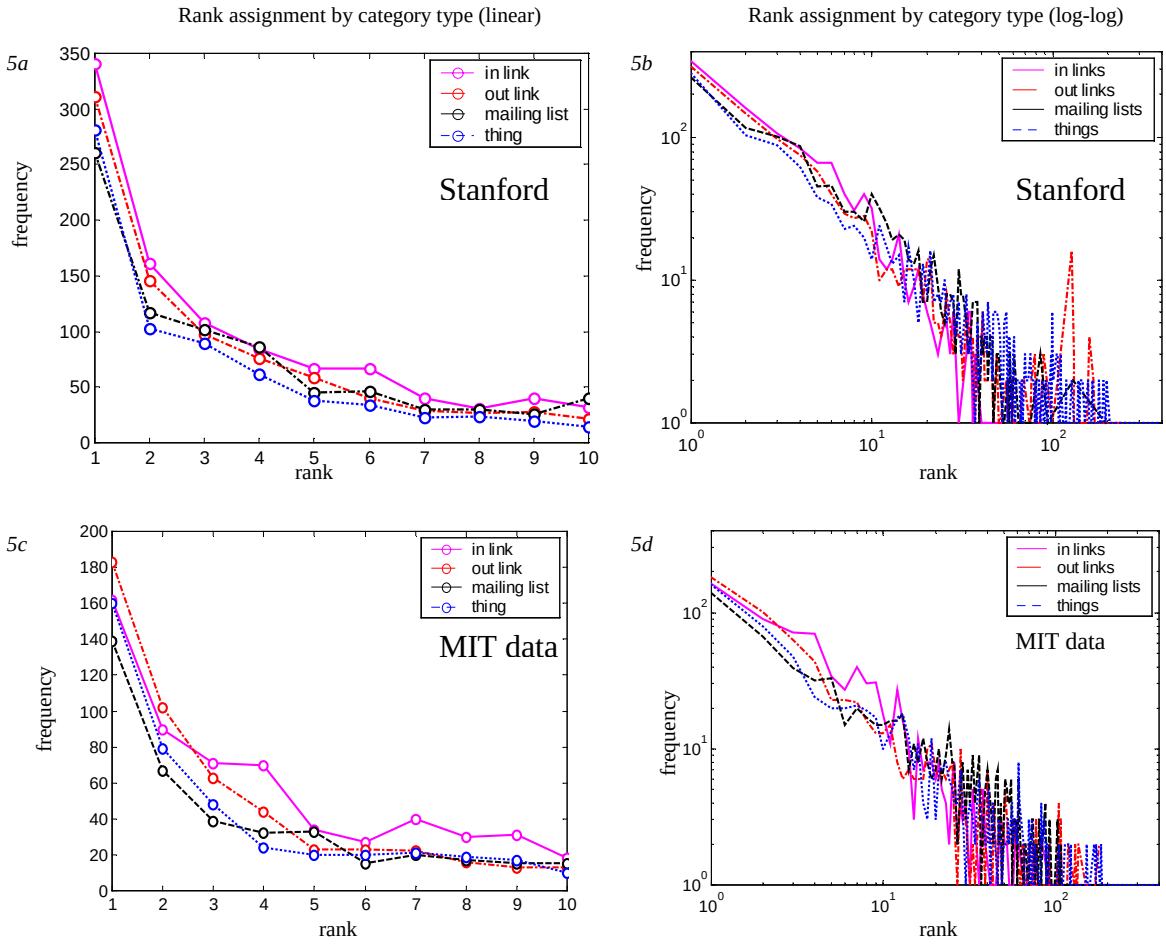

Figure 5a-d Figures 5a and 5c represent a linear scale plot showing how often we assigned each rank to a friend for the Stanford and MIT data respectively. Figures 5b and 5d are the log-log plot of the same data which illustrates the power law relationship.

Table 2 Top matches for a particular Stanford user, with the friends identified

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	Hands Christian Andersen
NO	2.41	Byung Joo Lee

The amount of data which could be used for ranking varied by type. For example, for Stanford the average number of terms, out links, in links, and mailing lists per user were 113, 22, 3, and 6 respectively. Note that the average numbers of terms, links, mailing lists, etc. a user has are not typical. This is due to the fact that they are distributed according to a power-law[2], meaning that most people have only a few items, but a few have a large number. Nevertheless the averages give a sense that people tend to include more text than links on their homepages. As a result, the fraction of friends ranked varied by the type of data used as shown in Table 3.

Since the number of terms recorded for a user was higher than the number of links, we were able to make more matches with respect to terms. However, the quality of matches based on terms 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 similarity value for which we would declare a match.

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 2 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[3]. Table 3 gives a summary of the results. We find that in-links are the most predictive followed by mailing lists and out-links, and finally text.

Table 3 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.

	Pairs Ranked		**Average rank**	
Method	**Stanford**	**MIT**	**Stanford**	**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

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 Figures 4b 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.

Finally, one may expect that friends should have the most in common, while friends of friends should have less in common (and so on). We see that this is indeed the case as shown in Figure 6. In this Figure we plot the average combined likeness score versus distance, taking into account text, links, and mailing lists. In line with our hypothesis, the result appears as a rapidly decaying function in which the likeness score quickly falls off as distance increases.

4. INDIVIDUAL LINKS, TERMS, AND MAILING LISTS AS PREDICTORS

Until now we have referred to shared items as an abstract concept. While the predictive algorithm simply takes into account the frequency of these items it is valuable to understand the types of items that contribute heavily to the prediction scheme. Intuitively one would expect some items to be shared only by friends, while others could be associated with almost anyone.

For this analysis, we attempted to measure an individual item's ability to predict whether two people who mention it will link to one another. The metric used was 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)/2, where N is the number of users associated with the item.

Table 4a-d lists the top 10 ranked terms, (in and out) links, and mailing lists as ranked by the equation above for each of Stanford and MIT. What we find is that shared items that are unique to a community are pulled to the top. Over general or popular terms such as “Electrical Engineering” are pulled further down.

While our technique appears to work quite well in representing key groups of individuals, some caution is necessary in over interpreting the broadness of these results as the measure favors smaller, tightly linked, groups. For example, the top phrase for MIT, “Union Chicana” appears in the home pages of five users. In this set five pairs of users have direct links between their pages. The ratio by our equation is therefore .5. Similarly, the last phrase “Russian House,” appears in five pairs among 14 users yielding a ratio of .055. However, what is interesting is that a different set of shared items is at the top of the Stanford and MIT lists. These differences are consistent and can be explained by real-life differences between the communities.

[3] 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.

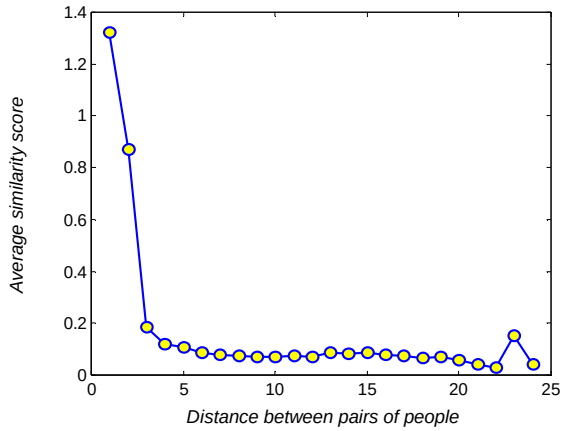

Figure 6 illustrates the relationship between the average likeness score and the number of hops between individuals.

For example, in the MIT list five of the top ten terms are names of fraternities or sororities. In the Stanford list only KDPhi, a sorority, appears in the list. This is consistent with the residential situation in the two schools. In addition to its dormitories, MIT has over thirty living groups (fraternities, sororities, and co-ed). Nearly 50% of all undergraduate males reside in one of these living groups for a full four years. Even students who choose to live in a dormitory tend to stay in the same one for all four years. In contrast, at Stanford only 9 of the 78 undergraduate houses are fraternities and sororities. Students not living in a fraternity or sorority reenter the housing lottery every year and may change their place of residence. Residential choice is a much less integral part of Stanford student life and is much less likely to appear on a Stanford student homepage.

Recall that a shared in-link is a page that points at two individuals (which link to each other). In both the Stanford and MIT data this list is dominated by individual homepages. These homepages link to the person's friends, and these friends in turn link to one another, exposing a social clique. In other words friends have friends in common. Nine of the top ten for Stanford, and ten of the top ten for MIT are homepages for individuals.

Another notable difference between the sets of shared items is the strong prevalence of religious groups for MIT users[4]. Stanford on the other hand is much more varied in this category.

In both Stanford and MIT the metric shows consistent results are in which items are poor predictors. Frequently occurring terms such as large US cities, and degree titles (BA, MS, etc) dominate the bottom of the term lists. This is consistent with traditional homepage structure in which the users list their city of origin and their current degree aim ("I'm from Chicago and I'm getting my BS in Computer Science").

Poor links for both Stanford and MIT are also similar. Pointers to popular sites such as Yahoo and AltaVista do not provide useful predictive power. General institutional web sites such as www.stanford.edu for Stanford and www.mit.edu for MIT are also poor predictors.

For MIT and Stanford, the mailing lists that appear to be bad predictors fall into three main categories: very general discussion lists, announcement lists, and social activities.

While these results are by no means definitive in providing an understanding of the social working of two communities it is reassuring to find that they follow some intuition and match some real-world analogue.

5. FUTURE WORK

In limited experiments students presented with their best matches given by our algorithm frequently recognized the individuals listed, even if they had not expressly put a link to them from their homepage.

Individuals interact with many people on a regular basis, but do not link to all of them through web pages. The fact that we do not have this complete list of friends results in many false negative matches. 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). To reconcile this, a future direction for this work would go beyond homepages to obtain social links directly from users.

Additionally, while we have select four data sources in particular there are many others that can be used. For example, demographic information such as address, major, and year in school, may provide us with extra clues. These sources are also available and can be integrated into our automated techniques.

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. We have also characterized specific types of items from each of these categories which turn out to be good or bad predictors. Furthermore, because predictors vary between communities, we were able to infer characteristics of the communities themselves.

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

[4] The names of these lists have been blocked for MIT as mailing lists are not publicly available.

Table 4a-d The top items as measured by the ratio of linked pairs of users associated with the item divided by the total possible number of pairs. Each sub-table lists the top ten items for Stanford and MIT.

	MIT	Stanford
Top Phrases	Union Chicana (student group)	NTUA (National Technical University of Athens)
	Phi Beta Epsilon (fraternity)	Project Aiyme (mentoring Asian American 8th graders)
	Bhangra (traditional dance, practiced within a club at MIT)	pearl tea (popular drink among members of a sorority)
	neurosci (appears to be the journal Neuroscience)	clarpic (section of marching band)
	Phi Sigma Kappa (fraternity)	KDPhi (Sorority)
	PBE (fraternity)	technology systems (computer networking services)
	Chi Phi (fraternity)	UCAA (Undergraduate Asian American Association)
	Alpha Chi Omega (sorority)	infectious diseases (research interest)
	Stuyvesant High School	viruses (research interest)
	Russian House (living group)	home church (Religious phrase)
Top Out-links	MIT Campus Crusade for Christ*	alpha Kappa Delta Phi (Sorority)*
	The Church of Latter Day Saints	National Technical University Athens
	The Review of Particle Physics	Ackerly Lab (biology)*
	New House 4 (dorm floor, home page)*	Hellenic Association*
	MIT Pagan Student Group*	Iranian Cultural Association*
	Web Communication Services*	Mendicants (a cappella group)*
	Tzalmir (role playing game)*	Phi_Kappa_Psi (fraternity)*
	Russian house (living group) comedy team *	Magnetic Resonance Systems Research Lab*
	Sigma Chi (fraternity)*	Applications assistance group*
	La Unión Chicana por Aztlán	ITSS instructional programs*
Top In-links	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Individual's list of friends*
	Individual's list of friends*	Sorority member list*
Top Mailing lists	Summer social events for residents of specific dorm floor	Kairos97 (dorm)
	Religious group	mendicant-members (a cappella group)
	Religious group	Cedro96 (dorm summer mailing list)
	Religious group	first-years (first year economics doctoral students)
	Intramural sports team from a specific dorm	local-mendicant-alumni (local a cappella group alumni)
	Summer social events for residents of specific dorm floor	john-15v13 (Fellowship of Christ class of 1999)
	Religious a cappella group	stanford-hungarians (Hungarian students)
	Intramural sports team from a specific dorm	serra95-96 (dorm)
	"…discussion of MIT life and administration."	metricom-users (employees who use metricom)
	Religious group	science-bus (science education program organized by engineering students)

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 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.

7. ACKNOWLEDGEMENTS

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

8. 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] 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.

[5] P. Erods, and A. Renyi, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17; B. Bollobas, *Random Graphs* (Academic Press, London, 1985).

[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] HomePageSearch http://hpsearch.uni-trier.de/hp/

[9] InXight ThingFinder product page, http://www.inxight.com/products_wb/tf_server/index.html.

[10] 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.

[11] S. Milgram, ``The small world problem," *Psychology Today* 1, 61 (1967).

[12] J. Sharkes, M. Langheinrich, and O. Etzioni, Dynamic Reference Sifting: a Case Study in the Homepage Domain, Proceedings of the Sixth International World Wide Web Conference, pp.189-200 (1997).

[13] D. Watts and S. Strogatz, ``Collective dynamics of small-world networks," *Nature* 393, 440 (1998).

[14] Patricia Wallace, The Psychology of the Internet, Cambridge Univeristy Press, Cambridge, 1999