



Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Social Networks 25 (2003) 211–230

**SOCIAL
NETWORKS**

www.elsevier.com/locate/socnet

Friends and neighbors on the Web

Lada A. Adamic*, Eytan Adar

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

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 and tools to mine this information in order to extract social networks and the exogenous factors underlying the networks' structure. In an analysis of two data sets, from Stanford University and the Massachusetts Institute of Technology (MIT), we show that some factors are better indicators of social connections than others, and that these indicators vary between user populations. Our techniques provide potential applications in automatically inferring real world connections and discovering, labeling, and characterizing communities.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Homepage analysis; 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 (Larson, 1996; Gibson et al., 1998; Flake et al., 2000) 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 mentioning 'golf' and linking to other pages on the same topic would be part of the golf page collection. Such a page is not necessarily a homepage or even associated with a particular individual. In contrast, our

* Corresponding author.

E-mail addresses: ladamic@hpl.hp.com (L.A. Adamic), eytan@hpl.hp.com (E. Adar).

work focuses on *individuals'* homepages and the connections between them. By tapping into these digital representations for users we can learn a great deal about both virtual and real world communities of people.

Although homepage identification has been researched as a separate problem (Sharkes, 1997; Hoff, 1998), our analysis goes beyond identifying homepages to study the connections between them. Because pages that share a topic are likely to link to one another, and because the pages in our analysis can serve as proxies for individuals, we can use these connections to characterize relationships between people. Applying our techniques to this data we can begin to answer many questions. For example, are people who mention 'dance troupe' likely to link to each other? And furthermore, which terms are most indicative of connections: is 'dance troupe' a better indicator 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 entire 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 (Diaz, 2000). 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 an increased load on base stations relative to roads with no traffic.

Just as individual cell phone use produces road traffic data, the information side effect of writing homepages is the insight into social network structure it reveals. 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 mail, phone or live surveys. 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. For example, once co-authorship data was compiled electronically, scientific collaboration networks could be analyzed on a large scale (Batagelj and Mrvar, 2000; Newman, 2001). Similarly, tracking email interactions can provide an interesting birds-eye view of social networks (Schwartz and Wood, 1993). Extending this approach to homepage data presents an unprecedented opportunity to discover new and interesting social and cultural phenomena.

The data we study, as described below and in Fig. 1, comes from the following four different sources (although there are many more):

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.

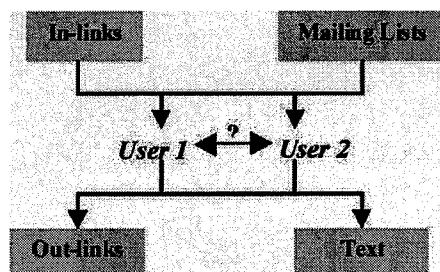


Fig. 1. There are four sources of information for a user: in-links and mailing lists which were provided by external sources, and out-links and text which were provided by the users themselves. All four can be used as a means of inferring relationships between the users.

1. *Text* on user's homepage provides semantic insight into the content of a user's page. Co-occurrence of text between users who link to each other usually indicates a common interest. We use multi-word "things" such as organization names, noun phrases, etc. instead of single words.
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.

We were interested in evaluating the ability of each of the above four sources of information to indicate 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 indicators, it was necessary to build a constrained data set. We achieved this by crawling the homepages 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 a web interface for exploring the social networks. Section 4 describes prediction schemes for link structures based on the information sources described above, and in Section 5 we discuss which particular types of information are indicative of social connections in different communities. In Sections 6 and 7 we provide areas for future work, potential applications of this technique, and draw general conclusions.

2. Homepage link structure

Milgram (1967) showed that real world networks are characterized by the small world phenomenon, where any two people in the world are connected through a short chain of

acquaintances. We looked for the same characteristics in the network of homepages. The fact that makes Milgram's original result surprising is that most people tend to move in close social circles tied to a geographic location, profession, or activity.

Watts and Strogatz (1998) incorporated both clustering, or the tendency of people to interact in groups, as well as small average geodesic distance into their small world network model. 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 (WWW) is also a small world network (Adamic, 1999; Albert, 1999). 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 MIT.

Homepage networks arise because it is popular for students to mention their friends on their homepages (Wallace, 1999), and link to those friends' homepages if they exist. They might be imitating lists they have 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. In limiting ourselves to university homepages, we found that relationships implied in links between pages were similarly limited (friend, colleague, student, dorm-mate, etc.). We will later describe ways in which our tools can be extended to deal with the broader set of relationships that links may imply in the Web at large.

For this study, we looked at all users having a homepage under the domains <http://www.stanford.edu> and web.www.mit.edu. These 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.

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 omitted homepages without links and 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. Among both Stanford and MIT users who link to at least one homepage, over 50% of the links are reciprocated. 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

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

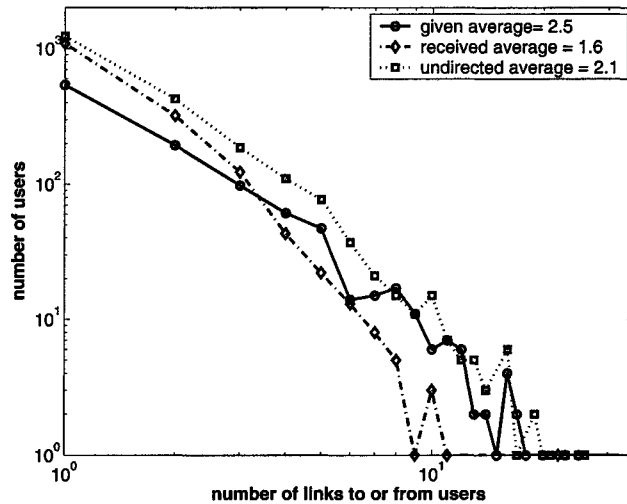


Fig. 2. Distribution of given, received, and undirected links in the Stanford social web. Note the log–log scale. The averages were 2.5, 1.6, and 2.2 for given, received, and undirected links, respectively.

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. Outside of constrained data sets it may be more difficult to easily detect such occurrences. In these cases link directionality may play a more important role and should be maintained.

Fig. 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 to users. Some users are very popular, attracting many links, while most get only one or two. The inequality within indegree and outdegree distributions is a characteristic the homepage network shares with 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. In contrast, in a random graph any two people are equally likely to be connected, which implies a Poisson degree distribution (Erdős and Renyi, 1960; Bollobas, 1985).

Next we computed the sizes of the connected components in the network. Even though the average number of links given by a user is only 2.5, a giant virtual connected social network of 1265 homepages exists. It accounts for 58% of the users, while a few smaller networks make 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.

Fig. 3a and b shows 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 geodesic distance between pairs of users is a mere 9.2 hops following links on homepages. Comparing Fig. 3a

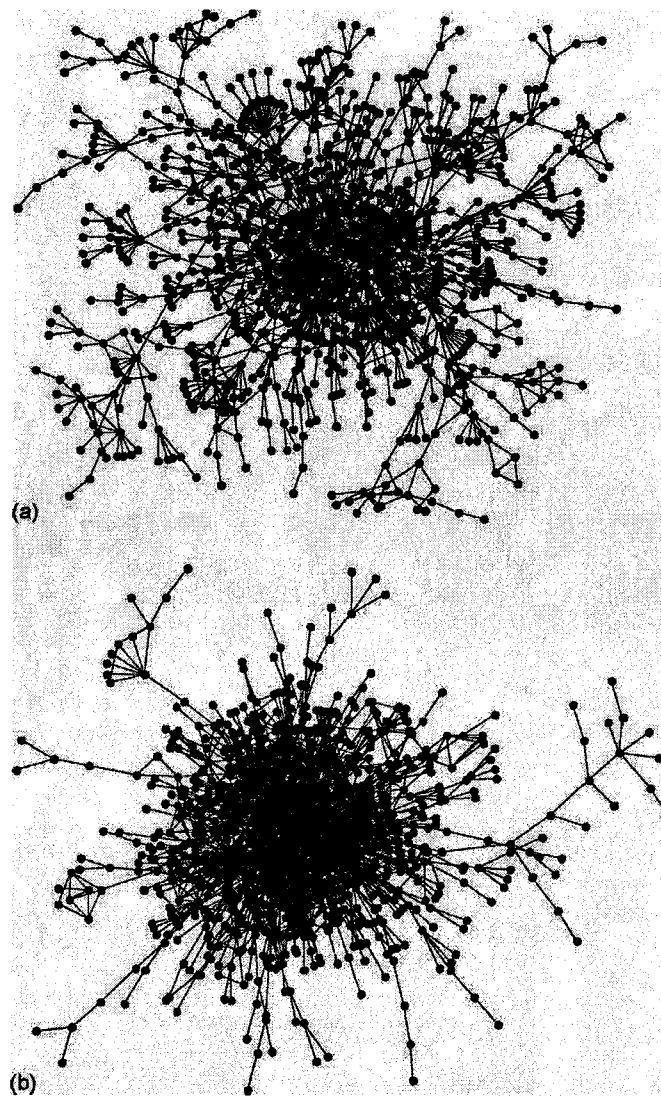


Fig. 3. Graph layouts of the (a) Stanford and (b) MIT social webs, with points representing homepages, and lines representing the links between them.

and b we see that MIT appears as a more tightly knit community. Indeed, this is reflected in the lower average geodesic of 6.4.²

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

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.

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 is 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.

3. Homepage analysis tools

While the bare link structure of homepage communities lends itself to interesting network analysis, it does not necessarily provide us with an understanding of why these links exist. We developed a web interface that allows one to interactively navigate through the social network while finding context for the social ties from the additional data gathered from the homepages. Specifically, the interface allows one to:

- (A) *search for individuals* with homepages by searching for names or browsing a directory;
- (B) *view user details* such as text and links in a homepage;
- (C) *visualize* the graph of a user's social network in both a local and a global context;
- (D) *explore a user's personal social network* by listing whom the user links to and who links to them, and seeing what those users have in common;
- (E) *match* a specific user to others using links, text, and mailing lists using an algorithm described below.

A demonstration of this application for the Stanford community is available at <http://www.hpl.hp.com/shl/papers/fnn/>. The majority of the tools provided are implemented as web scripts or java applets. Each feature of the system is elaborated on in the following sections.

3.1. Search and navigation

Our tools provide search and navigation of the social network we constructed. One can either type the username or part of the name of any individual in our database and be routed to their information page. Navigation is provided through links found both on directory pages and each information page (see Fig. 5, for example).

3.2. User details

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. InXight Software's ThingFinder³ 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

³ <http://www.inxight.com/>.

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 that 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. The tool presents categorized terms, thus giving an overview of homepage content that may have been fragmented over many pages and time consuming to navigate.

In-links to Stanford homepages were collected by querying Google. Within the MIT websites multiple URLs correspond to the same page, requiring us to use AltaVista’s wildcard search to gather in-links for MIT. Finally, complete lists of subscribers to mailing lists were obtained from the main mailing list server of each institution. At the time of our study, information on 95% of the lists at Stanford was publicly accessible from outside of Stanford. The remaining 5% of the mailing lists were excluded from our study. All information about the MIT mailing lists is internal to MIT. Because of this, and because users have some expectation about the privacy of their e-mail subscriptions, our public tools do not display list names for either Stanford or MIT.

Fig. 4 provides an example of the output generated by the user description page for one specific Stanford user. One of the side effects of our extraction techniques is the lack of contextual information in the overview pages. Our system presents bare terms and links rather than entire sentences. An advantage of such a format is that it can summarize multiple pages into a concise and categorized form. However, it is not possible to determine from this view why the user chose specific terms. For example, we might have extracted the term ‘Britney Spears’ from a homepage. At first glance one may think that the person is a fan of the pop star, when they could have in fact written ‘I hate Britney Spears’. Although people usually write about things they are involved in and appreciate, one should keep in mind that a fair fraction of the time they mention things they do not like and are not associated with. Future improvements to our system will provide more contextual cues (perhaps linking to the sentence in which a term appeared).

3.3. Visualization

Graphical representation of the network is generated by a java applet that lays out the largest connected component of the graph. Fig. 3a and b are snapshots of the result. The applet interface allows one to locate specific users within the global graph. Users may also drag and move individual nodes in order to better see the link structure around an individual.

An alternative to the global view are the images generated for each user individually. Fig. 5 illustrates such an image. In it the user, as well as their first- and second-degree neighbors, are illustrated and labeled. This allows for viewing users in the context of their local network.

3.4. Exploration of a user’s personal social network

In network analysis one is interested in the connections between users. Our interface allows researchers to get details about web links and hopefully understand why they ex-

Things found on Sarah's page

CITIES: Stratford

PERSONS: adam fingerhut, faulkner, jeannie..., marcela..., mark ..., morrison, robert..., ryan...

NOUN GROUPS: academic excellence freshman, admission counselor, business enterprise trust, gown women, research assistant, undergraduate admission, business inquiries, computer databases, history class, implement undergraduate recruitment strategy, language instructor, transfer applications, yield program

MISC: admit weekend, assistant, b.a., chappell-lougee grant, cuernavaca, designer, east palo alto tennis, education, gpa, her own words, jarrid ..., marlon ..., masque, oxford, stanford admit weekend, student affairs, tutoring, vice provost, yahoo

ORGANIZATIONS: office of undergraduate admission, stanford english department, stanford office

POSITIONS: admission counselor, coordinator, editor, president

Out Links

[Two links to friend's pages]

<http://admit.stanford.edu/admit>

<http://www.altavista.digital.com/>

<http://www.stanford.edu/>

<http://www.yahoo.com/>

In Links

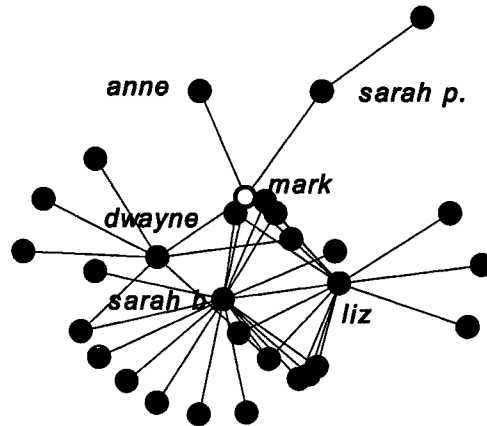
no in links

Sarah subscribes to 5 mailing lists (names are suppressed)

Fig. 4. Sample user data listed by category.

ist. Fig. 5 illustrates this facility for a Stanford student. Individuals that are linked to are listed in one section (for example, Dwayne) and those who link to the user are listed in another (Anne). Clicking on the name of an individual whom the user is connected to shows the individual's list of connections in turn. Thus one is able to browse the entire social network by following person-to-person links. Clicking on the "What do x and y

Mark



things and links found on Mark's web pages
people who have things in common with Mark

Here are the people whom Mark links to.
 Click on the name to find out who that person is linked to

Dwayne What do Mark and Dwayne have in common?
Sarah B What do Mark and Sarah have in common?

Here are the people who link to Mark.
 Click on the name to find out who that person is linked to

Anne What do Mark and Anne have in common?
Elizabeth What do Mark and Elizabeth have in common?
Sarah B What do Mark and Sarah have in common?
Sarah P. What do Mark and Sarah have in common?

Fig. 5. The link navigation interface allows one to visualize a user's neighborhood and follow homepage links.

User 1	User 2
Things In Common	
CITIES: escondido, athens, cambridge	
NOUN GROUPS: birth date, student association, undergraduate studies	
MISC: computer science, electrical engineering, ntua, ph.d., general lyceum, toefl, computer	
COUNTRIES: greece	
ORGANIZATIONS: national technical university of athens	
Pages pointed to by both users (out-links in common)	
http://www.stanford.edu	Stanford University Homepage
http://www.kathimerini.gr	Athens news
http://www.stanford.edu/group/hellas	Hellenic association
http://www.ntua.gr	National Technical University of Athens
http://ee.stanford.edu	Electrical Engineering Department
Pages pointing to both users (In-links)	
...	Dora's homepage
...	Dimitrios friends list
4 mailing lists in common	
greek-sports	sports list for Hellas members
hellenic	Hellenic association members
ee261-list	Fourier transform class list
ee376b	Information theory class list

Fig. 6. Items shared between two Stanford users who link to each others' homepages.

have in common" link produces a page listing the items shared by users x and y . Fig. 6 shows an example list of items shared by two users. From this view it is possible to see if two individuals share social activities, live in the same location, share academic interests, etc. This tool allows researchers to get a sense of the type of relationship that two users have.

3.5. Predicting relationships

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 associated with another, we rank all users by their similarity to that person. Our matchmaking algorithm is based on the well-established result in sociology that friends tend to be similar (Feld, 1981; Carley, 1991). Hence the more things two people have in common, the more likely they are to be friends, and the more likely they are to link to each other on their homepages.

Similarity is measured by analyzing text, links, and mailing lists. 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 five people mention the item, then its weight drops down to $1/\log(5)$ or 0.62. To summarize:

$$\text{similarity}(A, B) = \sum_{\text{shared items}} \frac{1}{\log[\text{frequency}(\text{shared item})]}$$

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.

We evaluate the performance of the algorithm by computing the similarity score for all users with respect to a single individual, and then ranking the users according to their similarity score. We expect individuals linking to each other on their homepages to be more similar to each other than randomly chosen pairs, and hence have higher rank. We measure this effect in two steps. First, we measure how many of the linked individuals can be ranked at all, that is what fraction has a non-zero similarity score. Second, we record what rank they were assigned.

Two users who are friends can falsely appear to have nothing common if one or both have very little information on their homepages. 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 would not be able to rank the friends with respect to each other based on out-links or text because there would be no overlap.

The amount of data available to our ranking algorithm 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 (Adamic and Adar, 2000), 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 linked pairs that could be ranked (shared at least one item) 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.

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

Linked	Likeness score	Person
Clifford		
No	8.25	Eric
Yes	3.96	John
No	3.27	Desiree
Yes	2.82	Stanley
No	2.66	Daniel
No	2.55	Wei
Yes	2.42	David
No	2.41	Hans
No	2.41	Byung

Table 3
Coverage and the ability to predict user-to-user links for four types of information about the user

Method	Pairs ranked		Average rank	
	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

The average rank was computed for matches above a threshold such that all four methods ranked an equal number of users.

In order to evaluate the success of our link 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 those the user indicated having a connection with by linking to their homepage. Table 2 shows an example of our procedure. We measured the success of our procedure in terms of the placement of homepage ties on the ranked list of matches based on similarity.⁴ 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.

Fig. 7a–d shows where those individuals directly linked to were placed. They were more than twice as likely to be ranked first than second, with the numbers decreasing from then on in a power-law fashion, as shown on the log-log plot in Fig. 7b and d. This means that the algorithm most frequently gives homepage ties a high rank, as one would hope, but every once in a while it gives an individual the user explicitly links to 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 in the homepage network as shown in Fig. 8. In this figure we plot the average combined likeness

⁴ The measure is asymmetric with respect to a pair of friends. Person A can rank as first for person B, but person B might only rank third for person A.

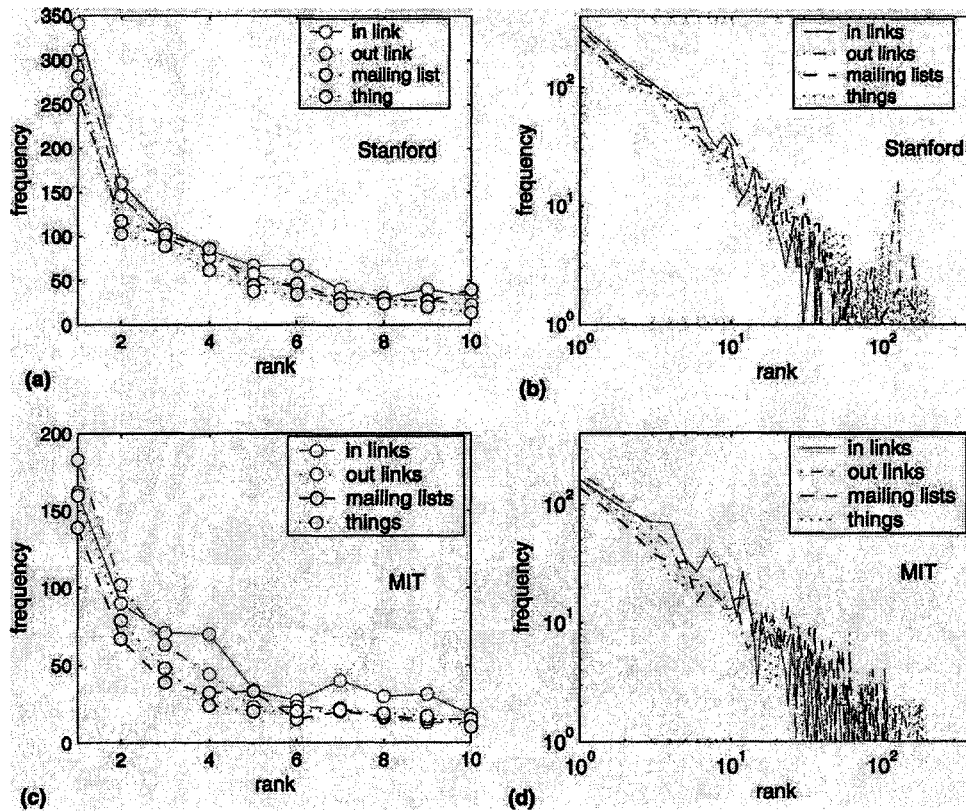


Fig. 7. Parts (a) and (c) represent a linear scale plot showing how often we assigned each rank to a friend for the Stanford and MIT data, respectively. Parts (b) and (d) are the log–log plot of the same data which illustrates the power–law relationship.

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 indicators of social connections

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 the cohesiveness of subgroups associated with a data item. 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.

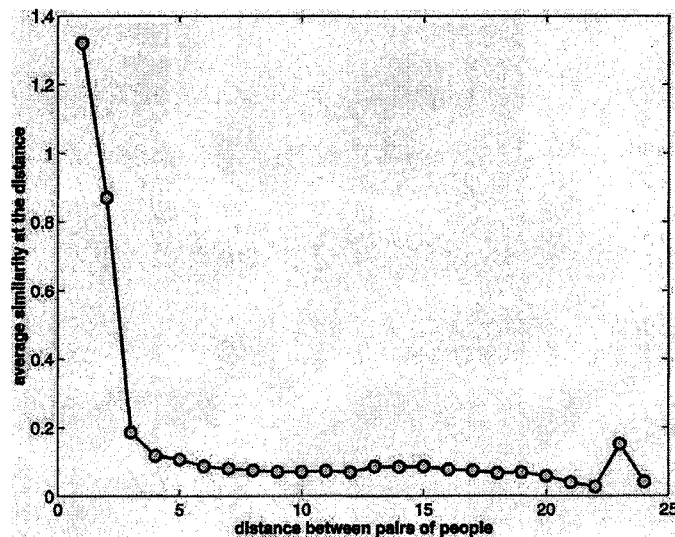


Fig. 8. Illustrates the relationship between the average likeness score and the number of hops between individuals.

Table 4 lists the top 10 ranked terms, (in and out) links, and mailing lists as ranked by density separately for Stanford and MIT. We find that shared items that are unique to a community are pulled to the top. Very general or common terms such as “Electrical Engineering” are pulled further down.

Note that the density measure favors smaller groups that are more likely to be tightly knit. Large groups would require individuals to link to a much larger number of others in the same group in order to achieve the same overall group density. For example, the top phrase for MIT, “Union Chicana” appears in the homepages of just five users. In this set five pairs of users have direct links between their pages, or each user has one link to someone else in the group on average. The density ratio by our equation is therefore 0.5. In contrast, the last phrase in the list, “Russian House,” appears in five pairs among 14 users yielding a density ratio of 0.055. In order to achieve the same density as the “Union Chicana” group, each user of the “Russian House” group would have to link to 6.5 other “Russian House” users on average, a condition which is much more difficult to meet. Both densities are significantly higher than the overall density of the entire network which is 0.0025. We can also verify that they are significant by using the normal approximation to the hypergeometric distribution (Wasserman and Faust, 1994, p. 272). We find the probability that 5 links are observed among 14 nodes by chance to be less than 10^{-5} .

Given the influence of size on group density, one must be careful not to over interpret numerical differences in densities between groups of varying sizes. What one can do, however, is to compare the qualitative differences in the most influential foci at Stanford and MIT. These differences are consistent and can be explained by real life differences between the communities.

For example, in the MIT list 5 of the top 10 terms are names of fraternities or sororities. In the Stanford list only one sorority, KDPHi, appears in the list. This is consistent with

Table 4

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

MIT	Stanford
(A) Top phrases	
Union Chicana (student group)	National Technical University of Athens (NTUA)
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)	Undergraduate Asian American Association (UCAA)
Alpha Chi Omega (sorority)	Infectious diseases (research interest)
Stuyvesant High School	Viruses (research interest)
Russian House (living group)	Home church (Religious phrase)
(B) 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, homepage)	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
(C) 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	Individual's list of friends
Individual's list of friends	Sorority member list
(D) 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-year (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)

Each part (A–D) lists the top 10 items for Stanford and MIT.

the residential situation in the two schools. In addition to its dormitories, MIT has over 30 living groups (fraternities, sororities, and co-ed). Nearly 50% of all undergraduate males reside in one of these living groups for a full 4 years. Even students who choose to live in a dormitory tend to stay in the same one for all 4 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 an in-link shared between homepages is a web page that points to both pages. 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. Nine of the top 10 for Stanford, and 10 of the top 10 for MIT are homepages for individuals. This is consistent with the fact that relationships develop more often and faster between people with mutual friends (e.g. Davis, 1967).

Another notable difference between the sets of shared items, mailing lists in particular, is the strong prevalence of religious groups for MIT users.⁵ Stanford on the other hand is more varied in this category.

Both the Stanford and MIT data sets are consistent in the types of items our metric evaluates to be poor predictors. Frequently occurring terms such as large US cities, and degree titles (B.A., M.S., 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 B.S. 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 websites such as <http://www.stanford.edu> for Stanford and <http://www.mit.edu> for MIT are also poor predictors.

For MIT and Stanford, the mailing lists that appear to be poor predictors (none or very few of the list subscribers link to one another) include announcement lists for large groups such as large classes or entire departments, popular activities such as the annual formal dance, movie or TV watching, and individual sports with large membership such as cycling, running and snowboarding.

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 observed real world properties of the two universities.

5. Discussion and future work

A notable characteristic of this study is that it uses only partial information (a fraction of individuals' social contacts and a fraction of their interests), and yet on the scale of an entire university we observe patterns such as small world properties of the networks and focal items around which groups of individuals cluster. The partial nature of the data introduces a number of biases. For example, homepage links can only exist between users with

⁵ The names of these lists have been blocked for MIT as mailing lists are not publicly available.

homepages, so that individuals without a homepage are absent from the study. Nevertheless, given a large enough sample, we find patterns consistent with the real world social structure of the universities.

The feasibility of using Web data in aggregate to extract information is also demonstrated in a study of Web translation tasks (Grefenstette, 1999). The study found that simply querying a web search engine with possible translations of a compound word or phrase frequently indicated the correct translation in the form of the number of pages corresponding to the phrase. For example, the compound German word “Apfelsaft” has two possible translations. The most common translation, “apple juice” has 92,500 hits in the Google search engine, while “apple sap” has only 202 hits. The study showed that treating occurrences of text statistically rendered the Web a useful tool for machine-based translation tasks, despite the numerous misspellings and poor grammar ubiquitous on web pages. Similarly, our analysis of homepages and their links in aggregate has yielded some interesting patterns despite the absence of many real world connections on-line.

Although much can be gleaned from the Web data alone, more traditional user studies can be used to supplement the Web data or confirm inferences drawn from it. 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. This is consistent with the fact that individuals do not list everybody they interact with on their homepages. This fact also causes our matching algorithm, which uses the Web data alone, to have 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 associations in our case). To reconcile this, a future direction for our work would go beyond homepages to obtain social links directly from users.

Additionally, while we have selected four particular data sources, 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. Some of these sources are also available online and can be integrated into our automated techniques.

Further, our study and the vast majority of social network studies are restricted to a small domain, such as a university, neighborhood, or organization. However, homepages are located on the World Wide Web, so that in principle homepage networks spanning the entire globe can be harvested and analyzed. With tools that can automatically identify homepages (Sharkes, 1997), this is not such a daunting task, and it presents an opportunity to study social networks on an unprecedented and global scale. Moving outside of a university setting will require some changes to the analysis and assumptions we have made. The number of possible relationships, for example, is enormous when looking at the Web at large. Individuals may no longer be friends or neighbors. Instead one individual may be a fan (one of thousands) of another, be a family member, or share an interest or hobby (without ever having met). In such an environment it would be useful to examine other pieces of information. For example, link directionality, the context of the link (the text around a link, for example), or the ratio of in-links to out-links would help to classify various relationships. By modifying our tools we believe that future work can include wider and more varied online communities.

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 can be an expensive and time-consuming process of conducting a series of mail, phone or live interviews. Studying social networks online can give us rich insight into how social bonds are created, but requires little more effort than running a crawler on homepages.

In this study we have demonstrated a means of leveraging text, mailing list, and in and out-link information to analyze network structure. We have also characterized specific types of items from each of these categories that act as good indicators (individuals associated with an item tend to link to each other) or bad indicators (items which are too general to be indicative of social connections). Furthermore, because indicators 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 labeled clusters of users. Thus, not only is it possible to find communities, but we can describe them in a non-obvious way.

Acknowledgements

The authors would like to thank Rajan Lukose, Bernardo Huberman, and T.J. Giuli for their valuable advice and comments.

References

- Adamic, L.A., 1999. The small world Web. In: *Proceedings of the European Conference on Digital Libraries*, Springer, Berlin.
- Adamic, L.A., Adar, E., 2000. Frequency of Friendship Predictors, <http://www.hpl.hp.com/shl/papers/fnn/>.
- Albert, R., et al., 1999. The diameter of the World Wide Web. *Nature* 401, 130.
- Batagelj, V., Mrvar, A., 2000. Some analyses of Erdős collaboration graph. *Social Networks* 22, 173–186.
- Bollobas, B., 1985. *Random Graphs*. Academic Press, London.
- Carley, K., 1991. A theory of group stability. *American Sociological Review* 56 (3), 331–354.
- Davis, J.A., 1967. Clustering and structural balance in graphs. *Human Relations* 20, 1181–1187.
- Diaz, S., 2000. Cell Phone Signals Touted to Fight Traffic Wars, San Jose Mercury News, <http://www.mercurycenter.com/svtech/news/indepth/docs/traf012100.htm>.
- Erdős, P., Renyi, A., 1960. On the evolution of random graphs. *Publications of Mathematical Institute of the Hungarian Academy of Science* 5, 17–61.
- Feld, S.L., 1981. The focused organization of social ties. *American Journal of Sociology* 86, 1015–1035.
- Flake, G., Lawrence, S., Lee Giles, C., 2000. Efficient identification of web communities. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, MA, pp. 150–160.
- Gibson, D., Kleinberg, J., Raghavan, P., 1998. Inferring Web communities from link topology. In: *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*.
- Grefenstette, G., 1999. The World Wide Web as a resource for example-based machine translation tasks. In: *ASLIB, Translating and the Computer*, vol. 21. London.

- Hoff, G., 1998. HomePageSearch, <http://www.hpsearch.uni-trier.de/hp/>.
- Larson, R.R., 1996. Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace, global complexity: information, chaos and control. In: Proceedings of the 1996 Annual Meeting of the American Society for Information Science. Baltimore, MD, USA.
- Milgram, S., 1967. The small world problem. *Psychology Today* 1, 61.
- Newman, M.E.J., 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98, 404–409.
- Schwartz, M.F., Wood, D.C.M., 1993. Discovering shared interests using graph analysis. *Communications of the ACM* 36 (8), 78–89.
- Sharkes, J., 1997. Dynamic reference sifting: a case study in the homepage domain. In: Proceedings of the Sixth International World Wide Web Conference, pp. 189–200.
- Wallace, P.M., 1999. *The Psychology of the Internet*. Cambridge University Press, Cambridge.
- Wasserman, S., Faust, K., 1994. *Social Network Analysis*. Cambridge University Press, New York.
- Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. *Nature* 393, 440.