Yes, There is a Correlation - From Social Networks to Personal Behavior on the Web

Parag Singla*
University of Washington
Seattle, WA, USA
parag@cs.washington.edu

Matthew Richardson
Microsoft Research
Redmond, WA, USA
mattri@microsoft.com

ABSTRACT

Characterizing the relationship that exists between a person's social group and his/her personal behavior has been a long standing goal of social network analysts. In this paper, we apply data mining techniques to study this relationship for a population of over 10 million people, by turning to online sources of data. The analysis reveals that people who chat with each other (using instant messaging) are more likely to share interests (their Web searches are the same or topically similar). The more time they spend talking, the stronger this relationship is. People who chat with each other are also more likely to share other personal characteristics, such as their age and location (and, they are likely to be of opposite gender). Similar findings hold for people who do not necessarily talk to each other but do have a friend in common. Our analysis is based on a well-defined mathematical formulation of the problem, and is the largest such study we are aware of.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.4.3 [**Information Systems**]: Information Systems Applications—*Communications Applications*; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Experimentation, Measurement

Keywords

Social Networks, Instant Messaging, Search, Demographics

1. INTRODUCTION

There is a famous saying, "A man is known by the company he keeps." The relation between a person's social interactions and personal behavior has been the topic of study among sociologists for many years. A brief scan through journals such as *Social Networks* [7] and *The American Journal of Sociology* [1] shows that this relation is a question of interest. Among the research issues that people attempt to address are: given that two people are connected, are they similar to each other? How does their connection affect their personal behavior? How does their behavior vary based on the type of connection? In this paper, we tackle the first, in the context of the internet, which is to see whether people who talk to each other are more likely to be similar to each other.

Whether this relation exists is not just a sociological question. If it exists, characterizing it could have a large impact on many Internet applications. For example, search engines could personalize their results to match not only a person's stated interests, but also the interests inferred from the user's social network. Knowing a person's social network would allow one to infer what likes and dislikes a person may have, what advertisements they may be more likely to take note of, etc. It could also lead to more intelligent chat clients that, for example, recommend a new friend to join a chat based on the interests shared by the friends already chatting.

In order to analyze the relation between communication and personal behavior, we need two sources of data: (1) who communicates with whom, and (2) the characteristics of each person in the communication network. For the first, we use an instant messaging network, and for the second, we use data from people's search history and their demographics.

Instant messaging (IM) has gained popularity in recent years, becoming a common form of communication for millions of people. One study, in 2005, found that 18% of Internet users use instant messaging daily [16]. Though email is still the primary communication medium, IM captures a different segment of communication. IM interactions tend to capture informal 'friends' connections between users, and thus represent an interesting social network in their own right.

For the characteristics of each person in the network, we turned to two sources. The first is the demographic data of the IM users, such as the person's age, gender, and geographical location. The second is based on personal interests. For this, we turned to Web search behavior. In the same study from 2005, it was found that 63% of Internet users visit search engines daily. A sampling of these searches demonstrates that users reveal personal interests and information through what they search for. For any interest that an Internet user has, it is very likely that he/she has at some point used a Web search engine to learn more about it. This makes the search engine query logs an ideal source of information about users' personal interests and behavior.

In this paper, we will show that there is indeed a very strong relation between who talks to whom on the instant messaging network, and what they search for. The correlation also holds for the category of their searches, their age, and their location. We found an anti-correlation between gender (that is, users who talk to each other are more likely to be of opposite gender than would be expected). We also found that these correlations strengthen with the total amount of time the two users spend talking. Interestingly, the correlation decreases with the amount of time spent *per message*; users who send very brief messages (perhaps indicating that they

*work done while at Microsoft Research

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

are closer friends, and thus need less formality in their communication) are more likely to be similar to each other. We also found that the more time a user spends per message, the more likely it is that he/she is talking to someone of the opposite gender. We also present additional studies, such as what happens for a pair of users that is not directly connected but does have a friend in common, and what happens if we condition on two users sharing some demographics, such as their location.

Our research falls in the domain of social network analysis [20] which has been an important area of study, with its primary goal being to understand the influences that nodes in a network have on their neighbors and how these influences propagate through the network. To the best of our knowledge, the particular problem that we tackle in this paper has not been studied before for such a large network. We define the problem in precise mathematical terms and then present a formalism which has the advantage of being very simple to understand and yet presents many insights into the data. Though our experiments have been done in the context of IM and search, we expect the same results would hold for a wide variety of networks, such as online gaming systems, newsgroups, and social web sites, and also for a wide variety of behaviors, such as what Web sites users visit, where they shop, etc.

The outline of the paper is as follows. We first provide theoretical motivation for the problem. The next section describes the datasets in detail. This is followed by the experimental set up and a wide variety of results. The paper concludes with some directions for future work.

2. THEORY

Consider a set of users in an online environment. Each user is represented by some relevant set of attributes. For example, the attributes of interest could be keywords searched, age, zip, gender etc. Let this set of users be denoted by U. Further, let some of the users interact with each other in some online interaction environment e.g. an instant messaging network. Let R be the relation denoting which pair of users interact with each other in the given environment. Further, let us assume that R also encodes the parameters associated with the kind of relationship which exists between each pair of users. For example, in the IM environment, these parameters could be the total talk duration, number of chat sessions etc. Given this setting, there is some underlying model $\Phi(U, R)$ which describes the distribution over various users (represented in terms of the attributes representing them) and the relationship between them. Let the corresponding distribution be denoted by $P_\Phi(U, R)$. One of the quantities of interest is $P_\Phi(U|R)$ i.e. what user characteristics hold given a particular instance of R. For example, one might want to predict the demographics of a set of users given that they talk to each other. Another quantity of interest is $P_\Phi(R|U)$ i.e. what relations hold given the set of user characteristics. An example of this would be predicting who talks to whom given the keyword searches performed by the various users in the domain.

Given the set of user characteristics U, a pairwise similarity vector can be calculated for each pair of users using some underlying similarity metric for each attribute. Let S_U denote the set of pairwise similarities which exist amongst users in set U. Now, let us make a simplifying assumption about the model. We will assume that the relationships between users depend only on the similarities between them, and not on the individual user characteristics. Mathematically,

$$P_\Phi(R|U) = P_\Phi(R|S_U) \tag{1}$$

For the simplicity of notation, we will denote S_U by S unless otherwise needed. Now, applying Bayes' rule, we get

$$P_\Phi(R|S) = \frac{P_\Phi(S|R)P_\Phi(R)}{P_\Phi(S)} \tag{2}$$

Let $S = \{S_{A_1}, S_{A_2} \cdots S_{A_m}\}$ be represented as the set of similarities for each attribute, S_{A_k} being the similarity for attribute A_k, $1 \leq k \leq m$. Equation 2 can then be rewritten as

$$P_\Phi(R|S) = \frac{P_\Phi(S_{A_1}, S_{A_2}, \cdots S_{A_m}|R)P_\Phi(R)}{P_\Phi(S)} \tag{3}$$

One common way to model a distribution such as this is by making a naive Bayes [5] assumption, which renders the attribute similarities independent of each other given R. This leads to the following equation:

$$P_\Phi(R|S) = \frac{P_\Phi(S_{A_1}|R)P_\Phi(S_{A_2}|R) \cdots P_\Phi(S_{A_m}|R)|R)P_\Phi(R)}{P_\Phi(S)} \tag{4}$$

The goal in this paper is to directly evaluate $P_\Phi(S_{A_k}|R)$ for all A_k. We should mention here that in evaluating these probabilities, our intention is not to calculate $P_\Phi(R|S)$. Rather, our goal is to analyze and understand these quantities by themselves. In particular, we would like to compare the conditional probabilities $P_\Phi(S_{A_k}|R)$ with their prior probabilities $P_\Phi(S_{A_k})$ to understand how the probabilities change when information about the relationship R is available.

3. DATASETS

We used two datasets for the analysis in this paper. The first corresponds to the interactions between different users on the MSN Messenger instant messaging network. This was obtained through the MSN Messenger logs maintained by Microsoft corporation. [1] The second piece of data corresponds to keyword searches made by various users on the Microsoft Web search engine (Windows Live Search), along with the information about personal characteristics such as the user's zipcode, age, and gender. Next we will describe these two datasets in detail.

3.1 Social Network Data

For the structure of the social network, we used data on user interactions in the MSN Messenger network, for a period of time in the summer of 2006. The raw data logged each event on the network, such as joining a chat session, chat invites, leaving a chat session etc., along with corresponding time-stamps. We obtained the raw data from Leskovec and Horvitz [14], which we processed to extract out relevant attributes. The final data contained 25 billion chat sessions, involving 162 million users. There were 3.3 billion pairs of users who interacted with each other at least once (edges in the social network), meaning on average a user pair interacted 8 times during the period of data collection.

The original data contains a wealth of information such as when each chat occurred, the length of each chat, etc. For this paper, we reduced this to a few statistics per user pair: the total number of chats between two users, the total number of messages exchanged, and the total time spent chatting. This means we are ignoring effects that may depend on, for example, the time of day, or the variance in chat session lengths (though we hope to examine these in

[1] It is important to point out here that we did not have any access to the actual contents of the chats that occurred on the network. The only information that we had access to was which user talks to which, and for how long / over how many sessions. Further, the information was available only in the form of anonymized user ids; there was no way for us to get back to original user identities using these ids.

Table 1: Aggregated Messenger Session

userid1	userid2	#sessions	#sent1	#sent2	duration

future work). Table 1 shows the fields stored for an aggregated session.

#sessions denotes the total number of individual sessions involved in the aggregate. *#sent1, #sent2* denote the total numbers of messages sent by each user aggregated over all the sessions. *Duration* is the total duration of all the sessions combined. As a side note, it is interesting to mention here that creating these aggregated sessions involved quite a few engineering skills. The sheer magnitude of the data made it unwieldy; aggregating pairs of user ids became an interesting task in and of itself.

3.2 Personal Interests Data

For information about people's interests and characteristics, we used a subset of Microsoft Web search data collected over a period in the summer of 2006.[2] The data contains half a billion searches performed by about 30 million distinct users[3]. So, on average, each user issued about 17 different searches during the period of data collection. The raw data contained more than 20 different attributes for each search log entry, including information about the user making searches (age, zipcode, gender etc.) and about the keywords being searched (query, query category etc.). As with the Messenger data, we reduced this to just 7 relevant fields, thus tossing out information about what time of day the search was issued, etc. We aggregated all searches performed by a given user into one entry, storing a concatenated list of all the search queries issued by each user during the given period. Table 2 shows what an aggregated search entry looks like. A *query/main-category/sub-category list* corresponds to a comma separated list of individual queries/main-categories/sub-categories respectively. *query* is the cleaned version of the keyword query issued[4]. Query category is decided based on classification of each possible keyword query into a two-level query type hierarchy. This hierarchy is pre-generated using the open directory project dmoz (http://dmoz.org) to classify various web pages. Each query is placed somewhere in the two-level hierarchy based on the documents it returns and where those documents lie in the original hierarchy. The first level of the hierarchy has 7 different *main categories* and includes things like software, entertainment etc. Each category in the first level is further divided giving rise to 35 different sub-categories. For example, software has sub-categories such as operating systems, programming etc. *Age group* is a discrete valued attribute representing 7 age-group buckets. For simplicity of notation, henceforth, we will simply refer to age group as age. *Gender* is either *male* or *female*. *Zip* is string of alphanumeric characters identifying a geographical location.

3.3 Joining the Data

Once we obtained the messenger data and the search data, they were joined together into one dataset where each tuple has the information about the aggregated messenger session as well as the searches for each user in the pair. This was simply done by scanning through the aggregated messenger data and appending to each tuple the aggregated search entries for the corresponding user ids from the search data. Only those sessions where search entries were available for both the users were kept. The resulting dataset consisted of 76 million tuples (one tuple for each user pair), corresponding to 13 million unique users. This joint tuple is shown in Table 3. Here *q* denotes the query list, *main-cat* denotes the list of main-categories and *sub-cat* denotes the list of sub-categories. *#messages* denotes the total number of messages exchanged between two users. This is the final form of data that we mined in the experiments described in the next section.

[2]Every effort was made to have maximum overlap between the collection times for the messenger data and the search data.

[3]This includes only those searches for which we had the user id information available

[4]Cleaning involves removing punctuation symbols, stemming, removing stopwords such as 'a', 'to','of' etc.

4. EXPERIMENTS

We performed a number of experiments on the joined messenger and search data described in the previous section. The first set of experiments establish a basic correlation between talking on messenger and similarity of various attributes. (That is, the conditional probabilities $P_\Phi(S_{A_k}|R)$ are significantly different from the prior probabilities $P_\Phi(S_{A_k})$, R being the messenger pair relation (see Section 2)). Further sets of experiments analyze how the correlation varies with varying the talk time, conditioning on certain attributes (such as zip) and the effect of having a neighbor in common rather than being directly connected on the messenger network. All our results are statistically significant ($p<0.01$), unless otherwise mentioned. We will first describe the way we compute the similarities for various attribute values. This will be followed by the details of our experiments.

4.1 Computing the Similarities

For the purpose of similarity calculation, we treated each attribute value as indivisible. Therefore, the similarity value is 1 if the attribute values are same, 0 otherwise. For queries, we also experimented with a softer similarity score which is explained later in this section.

Let $U = \{U_1, U_2, \cdots U_n\}$ be the set of unique user ids in our messenger/search environment. Let $A = \{A_1, A_2 \cdots A_m\}$ denote the set of attributes associated with each user. Each $A_k, 1 \leq k \leq m$, takes values from a finite domain D_k. For example, if A_k was gender, then $D_k = \{\text{male,female}\}$. In our case, A_k varies over queries issued, main-categories, sub-categories, zip, age and gender. Queries issued, main-categories and sub-categories will be referred to as *query attributes* as their value depends on the query issued by the user. Other attributes i.e. zip, age and gender will be referred to as *personal attributes*. We will use the notation $U_i.A_k$ to denote the value of k^{th} attribute associated with user U_i. For instance, if A_k was gender, $U_i.A_k$ would denote the value of the gender for user U_k. Further, U_{ij} will denote the user pair (U_i, U_j). Note that in case of multi-valued attributes e.g. when A_k is queries issued, $U_i.A_k$ will denote a multiset of values coming from the domain of attribute A_k.

Let us assume that we have been given a subset of user pairs, denoted by S. Given S and some attribute A_k, we would like to compute the probability $P(U_i.A_k = U_j.A_k)$ where (U_i, U_j) is a randomly chosen user pair from S. We can broadly divide the set of attributes A into two categories for the purpose of this probability calculation.

- **Single-valued attributes:** The probability calculation is straightforward in this case. The required probability is simply the fraction of the total number of pairs which take the same value for the given attribute. The single-valued attributes are age, gender and zip.
- **Multi-valued attributes:** In this case, the probability is the average over the probabilities of an attribute value being the

Table 2: Aggregated Search Session

userid	query list	main-category list	sub-category list	age group	gender	zip

Table 3: Joined Tuple

userid1	userid2	q1	q2	main-cat1	main-cat2	sub-cat1	sub-cat2

age1	age2	gender1	gender2	zip1	zip2	#sessions	#messages	duration

same for each pair of users $U_{ij} \in S$. Given a pair U_{ij} and an attribute A_k, the probability that A_k takes on the same value for U_i and U_j (denoted by $P_{ij}(U_i.A_k = U_j.A_k)$) is calculated as follows. It is the fraction of entries in the cross product $U_i.A_k \times U_j.A_k$ which correspond to the same value for both U_i and U_j. For example, let A_k be the queries issued and $U_i.A_k$ ={"OS performance", "OS windows", "OS mac"} and $U_j.A_k$ ={"OS performance", "OS Microsoft, "OS apple"} then, $P_{ij}(U_i.A_k = U_j.A_k) = 1/9$.

In the methodology described above, each query is treated as an indivisible string. The probability of match is non-zero (and equals 1) *iff* two query strings match each other exactly. As described earlier, we additionally treated each query as a bag of words to achieve a softer similarity. The multiset of words appearing in all the queries issued by each user was computed [5]. For example, if a user issued the queries in the set {"Red Dog", "Smart Dog", "Bulldog"}, then the corresponding multiset of words would be {Red, Smart, Bulldog, Dog, Dog}. These multisets can then be plugged into the equation above to calculate the desired probabilities for word matches. We call this new attribute *word* and it is also one of the query attributes.

Note that for the case of multi-valued attributes, there are many ways to compute the similarities of attribute values. These similarities can then be folded into some kind of probability calculations. For example, one can take the ratio of the size of the intersection and the size of the union of the attribute values. Or the dot product between the two multisets can be taken. The primary motivation for the way we did it is was that it is in some sense closest to the notion of "What is the chance that randomly selected attribute values for the multisets $U_i.A_k$ and $U_j.A_k$ are actually the same?". Nevertheless, exploring other similarity measures is an interesting direction for future work.

4.2 Establishing the Correlation

The main goal of our basic experiment was to find out: if A talks to B and C is some other random user, then is A likely to be more similar to B than C? That is, whether having additional information that a pair of users talk to each other on the messenger network increases the likelihood of their searches as well as other personal characteristics being similar. The motivation for comparing with a random pair of users is as follows. Given any two users in the messenger environment, there is some prior chance that they will be similar to each other, on any given attribute value. (Though these prior probabilities will be different for different attributes.) For example, there is probably a good chance that the word "restaurant" appears in the queries of a randomly selected pair of users given the common usage of this word. Then our goal is simply to discover any additional signal for similarity given the fact that the pair of users talk to each other on messenger. Mathematically, we compared

[5]This computation is done on the fly in our implementation.

- **Baseline:** $P(U_i.A_k = U_j.A_k|(i,j) \in R)$ where R is the set of all possible user pairs
- **Messenger:** $P(U_i.A_k = U_j.A_k|(i,j) \in M)$ where M is set of users who talk to each other

We will simply refer to these probabilities as similarities. Figure 1 plots the results in the form of a histogram. The left graph shows the results for query attributes, and the right graph shows for personal attributes. Each attribute has two bars, the first one giving the similarity (in percentage) for a random pair and the second one for a messenger pair [6]. We also give the results in tabular form (Table 4) to present the exact similarities.

These results clearly show that having the additional information that a pair of users talk to each other on messenger increases the likelihood of their query attributes being same. There is a particularly high jump in the similarity for queries for a messenger pair. The similarity is almost zero for the case of random pairs, demonstrating that exact query matches are very unlikely in general. For messenger users, the similarity is seven times higher. This could mean the users share an interest, or could indicate they might be searching for what they are talking to each other about, verifying which is an important direction for future work.

For the personal attributes, zip similarity is very low for random pairs (1%), but quite high for messenger pairs (13%). This indicates that people who talk to each other are quite likely to be located in the same geographical area. The similarity for age also is higher for the messenger pairs, indicating that people tend to talk within the same age group more often than not.

The case of gender is interesting. The probability of gender being the same decreases (and goes below the unbiased coin flip probability) given the additional information that two people talk to each other. This indicates that people are more likely to talk to persons of opposite gender on the messenger network, an interesting finding from a sociological point of view.

4.3 Varying the Talk-time

Having established the basic correlation, the goal of our next set of experiments was to find out: if A talks to B more than A talks to C, then is A likely to be more similar to B than C? That is, to analyze the effect on the similarity of user attributes as the total duration of talk-time on messenger network is varied. We binned all the messenger pairs into five bins based on the total duration of the time they talked to each other. The distribution of messenger pairs is skewed towards having a low talk time. Therefore, instead of having equal duration bins, we had bins having same number of messenger pairs. The lowest bin number corresponds to the least talk time. For each bin, we then calculated the probability of various attributes being same as in the basic experiment. These were then plotted against the baseline. Mathematically, we compared

[6]Some of the bars such as the one for baseline query similarity are barely visible because they almost coincide with X-axis.

Table 4: Similarities (%) comparing random pairs and messenger pairs

	Word	Query	Main Category	Sub Category	Zip	Age Group	Gender
Baseline	0.51	0.09	15.26	6.23	0.81	34.40	51.67
Messenger	1.00	0.62	16.68	7.59	13.00	64.19	48.74

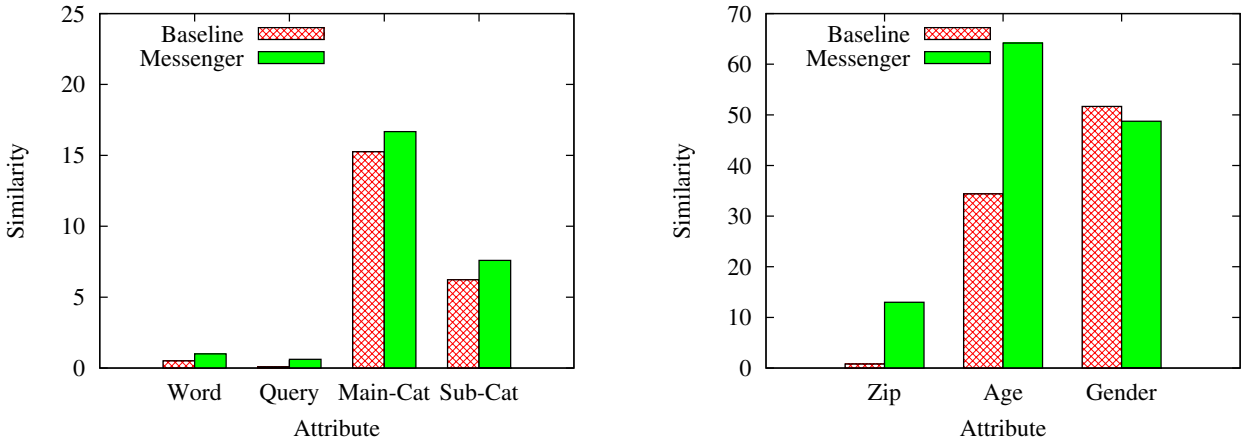

Figure 1: Similarities(%) comparing random pairs and messenger pairs: query attributes(left) and personal attributes(right)

- **Baseline:** $P(U_i.A_k = U_j.A_k|(i,j) \in R)$
- **Bin n** $(1 \leq n \leq 5)$: $P(U_i.A_k = U_j.A_k|(i,j) \in M, B_{ij} = n)$ where B_{ij} denotes the bin number for the talk duration of pair (U_i, U_j)

Figure 2 plots the histograms for various attribute similarities. On the left, we have query attributes and on the right, we have personal attributes. Baseline is represented by the first histogram bar. Subsequently, each bar represents a particular duration bin in the order of succession. Figure 3 shows the results in a graphical format. The X-axis represents the bins. There is a curve for each attribute showing its similarity value at each bin. This helps analyze the gradient in the similarity values as the talk duration is increased.

For any talk duration, the messenger pair similarity is more than the baseline similarity for all the query as well as personal attributes. Further, similarity increases monotonically with increasing talk duration for all the attributes. The only exception to this is gender where the similarity tends to fluctuate. These results lead to the conclusion that people who talk to each other more are more likely to be similar to each other. After the initial jump from the baseline, all the curves (except gender) rise more or less smoothly with increasing talk duration.

Further, we wanted to differentiate between users with many short sessions vs. few long sessions. Intuitively, this is like measuring the difference between superficial common friendships vs. deep friendships. This can be done by looking at how similarities vary as the average (instead of total) talk duration changes. As in the case of total duration, we used five bins dividing the average talk duration. The results in this case are qualitatively similar to the results for the case of total duration. As the average session length is increased, the similarities also increase monotonically.

Lastly, we wanted to analyze if anything can be inferred from the time taken to type each message (and including any time lag that happens before typing next message). A longer time for each message would probably indicate that either the participating users are not very interested in the conversation or they are being very careful with what they write or they are simply writing longer messages. As in the case of total and average talk durations, we used five bins dividing the average time spent per message. Figure 4 shows the histograms for the similarity values. Figure 5 plots the results graphically.

For all the query attributes and for zip, the similarities tend to decrease with increasing time spent per message. This is intuitive - shorter messages are probably indicative of a pair of users who are more familiar with each other (e.g., close friends), and such pairs are more likely to share interests. The last bin is an interesting exception, where similarities increase in some of the cases. We are not sure why this happens and analyzing whether there is a real trend there is a part of the future work. Age similarity does not seem to have any trend with increasing time spent per message.

For the case of gender, the similarity monotonically decreases with increasing time per message. In other words, users spend more time per message when they talk to people of opposite gender, a very interesting finding which warrants further exploration.

4.4 Conditioning on Personal Attributes

Above, we have shown that people who talk to each other tend to have similar interests, as evidenced by increased similarity in what they query for. We have also shown that they are more likely than random to be of similar age and location. One question that naturally arises is, is the similarity of interests due solely to the fact that people who talk to each other have similar demographics (for instance, if I and a friend are both in Seattle, we will both tend to query about local sporting events), or is there more to it, a genuine sharing of interests. To answer this, we compared the probability that queries (and their categories) were the same when one or more of the personal attributes (zip, age and gender) were the same. Mathematically, we compared

- **Baseline:** $P(U_i.A_k = U_j.A_k|(i,j) \in R)$
- **Conditioned Baseline:** $P(U_i.A_k = U_j.A_k|(i,j) \in R, C_{ij} = true)$ where C_{ij} is some Boolean condition specified on the attributes of U_i and U_j, e.g. C_{ij} could be $U_i.zip = U_j.zip$

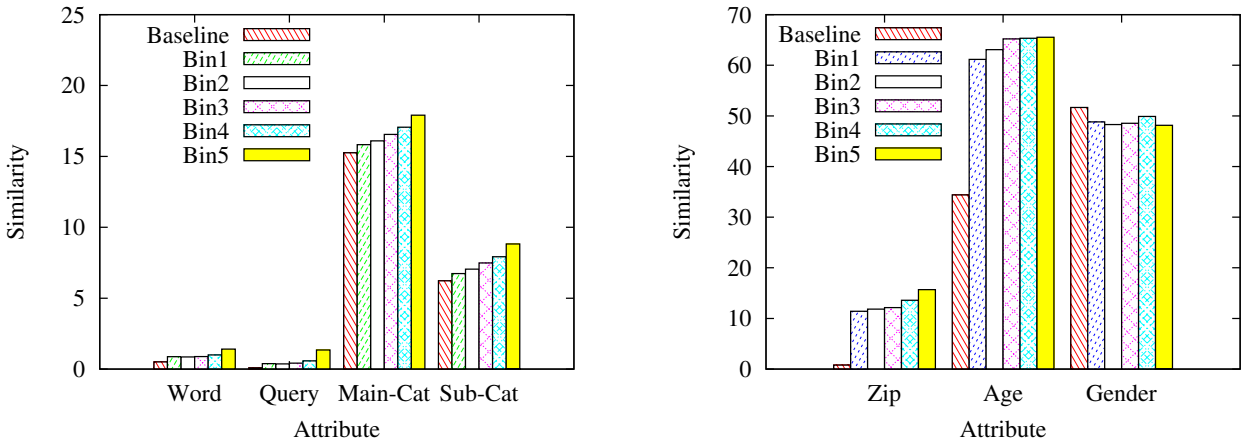

Figure 2: Variation in similarities(%) with total talk duration: query attributes(left) and personal attributes(right)

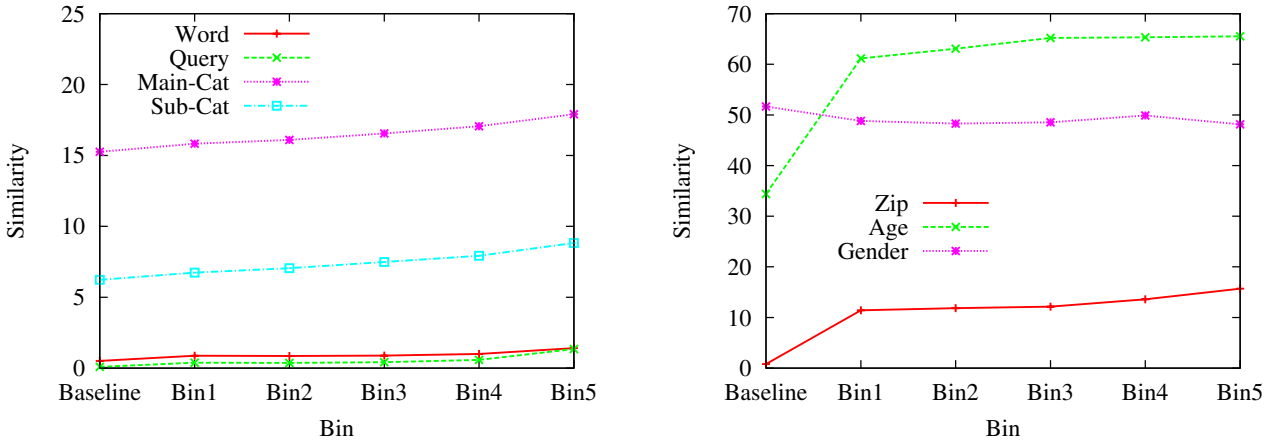

Figure 3: Variation in similarities(%) with total talk duration: query attributes(left) and personal attributes(right)

- **Messenger:** $P(U_i.A_k = U_j.A_k | (i,j) \in M)$
- **Conditioned Messenger:** $P(U_i.A_k = U_j.A_k | (i,j) \in M, C_{ij} = true)$ where C_{ij} is as before

Figure 6 shows the histograms for similarities conditioning on all personal attributes being same. First two bars are for the baselines (unconditioned and conditioned, respectively) [7] and next two bars for the messenger (unconditioned and conditioned, respectively). *C-Baseline* and *C-Messenger* are the shorthands for conditioned baseline and conditioned messenger, respectively.

The conditioned baseline similarities very closely follow the unconditioned baseline similarities. This is interesting because it says that the query similarity values for a random pair of users within same demographics are not much different from similarity values of a random pair of users. This is unlike messenger pairs, where similarities are higher for user pairs within the same demographics. More importantly, the conditioned messenger similarities are consistently significantly higher than the conditioned baseline similarities. This confirms our thesis that there is in fact a genuine sharing of interests between user pairs who talk to each other on the messenger.

[7]The difference between the two baselines was not statistically significant.

Figures 7, 8 and 9 show the results conditioning on zip, age and gender being same, respectively. The results are qualitatively similar to the case conditioning on all personal attributes being same.

4.5 Effect of Indirect Links

The goal of this final set of experiments was: If A talks to B and B talks to C, then what kind of similarity exists between A and C? In other words, we would like to find out whether users who have a friend in common also exhibit the same type of similarity as users who chat directly to each other. To answer this question, we simply need to calculate the similarities over a network where two users are connected if they have a common talking friend (by which we mean, if there is a person in common that each user talks to in the IM network.). We call such a network a 2-hop network as one has to traverse two hops in the original network to reach a user of interest. Analogously, the original messenger network can be called as the 1-hop network. (Note that a pair of users can belong to the 1-hop network as well the 2-hop network if they talk to each other directly and also have a common talking friend.) Since there are many more 2-way paths than 1-way paths, the size of 2-hop network would be much more than the size of the original network (in our case, it would have been about 10 times the size of the original network). Working on the complete 2-hop network would be too slow and inefficient. Therefore, we sampled the user pairs uniformly from

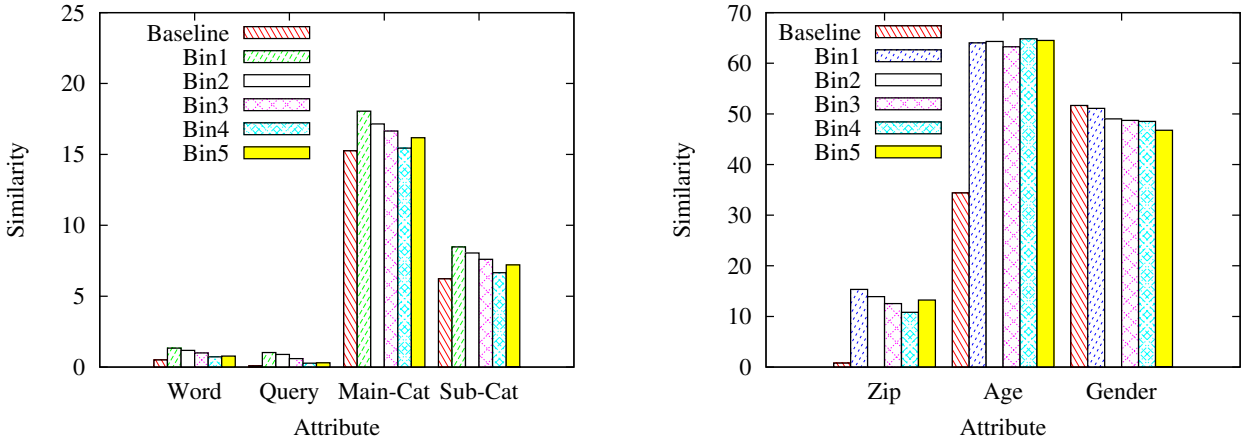

Figure 4: Variation in similarities(%) with average time spent per message: query attributes(left) and personal attributes(right)

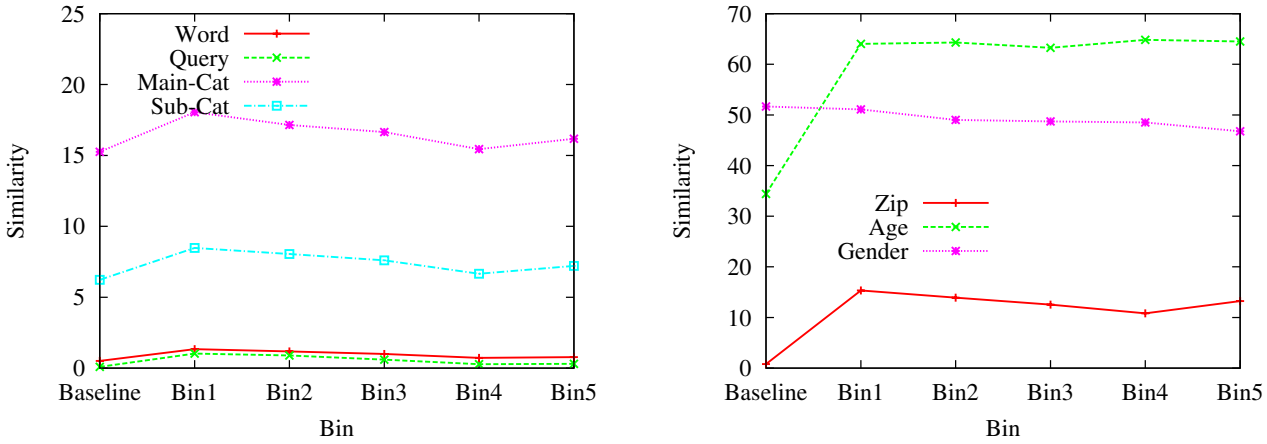

Figure 5: Variation in similarities(%) with average time spent per message: query attributes(left) and personal attributes(right)

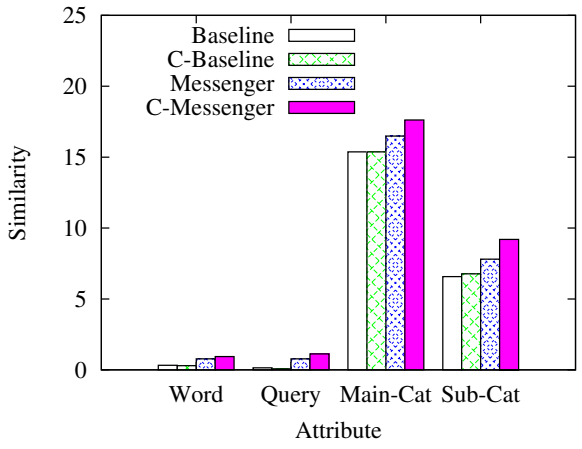

Figure 6: Similarities(%) of query attributes conditioning on all personal attributes being same

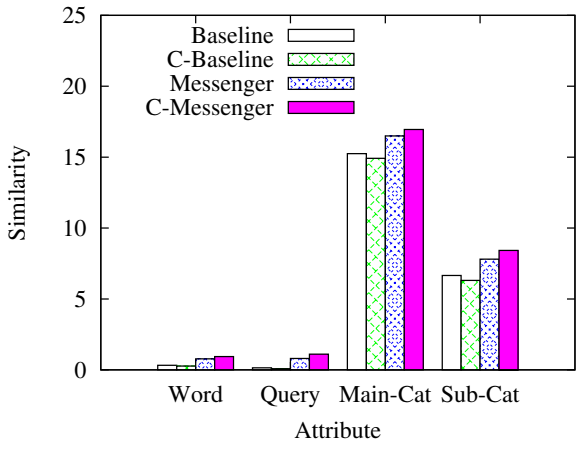

Figure 7: Similarities(%) of query attributes conditioning on zip being same

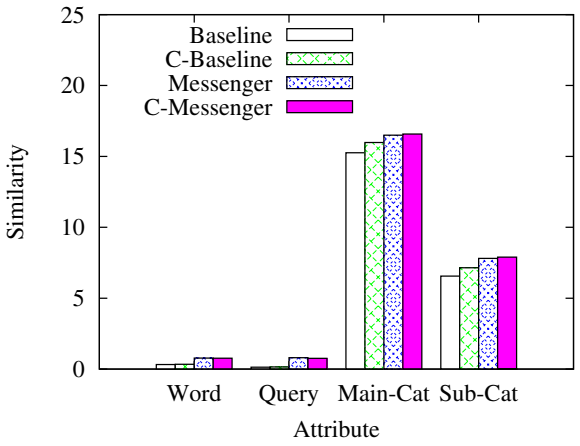

Figure 8: Similarities(%) of query attributes conditioning on age being same

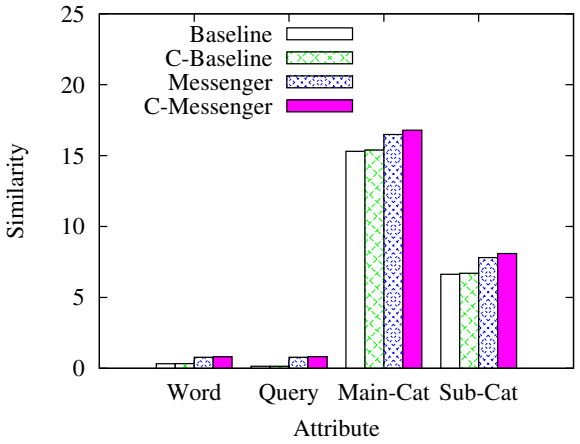

Figure 9: Similarities(%) of query attributes conditioning on gender being same

this network (the sampling was done online as we constructed the 2-hop network) to get a sample which was roughly the same size as that of the original 1-hop network. The results that we report are over this sampled subset of the 2-hop network [8]. Mathematically, we compared

- **Baseline:** $P(U_i.A_k = U_j.A_k|(i,j) \in R)$
- **1-hop:** $P(U_i.A_k = U_j.A_k|(i,j) \in M)$
- **2-hop:** $P(U_i.A_k = U_j.A_k|(i,j) \in M^2)$ where M^2 is the set of user pairs in the 2-hop network

Figure 10 shows the histograms for the similarities. As before, the left graph shows the similarities for query attributes and the right graph for the personal attributes. The first histogram bar represents the baseline, second the 2-hop network and the last one the 1-hop network. As one would expect, the similarity values in 2-hop network are somewhere in between the 1-hop similarities and random pair similarities.

[8]Although, we report the results over a particular sample, we tried few iterations of sampling and the results were invariant for all practical purposes.

For all the query attributes and for zip, the 2-hop similarities are midway between the 1-hop and baseline similarities. For age, the 2-hop similarity is very close to the 1-hop similarity.

At 2-hops, it is much more likely to find people with the same gender. This makes sense given that people of opposite gender are more likely to talk to each other in the original network. Since friends are more likely to be of opposite genders, friends of friends are more likely to be of same gender.

Overall, one can conclude from these results that, people who have a common talking friend are more likely to be similar than a random pair of users.

Now, generalizing the above idea, one might ask the question "What kind of similarities exist between users who are connected to each other through a chain (of some length) of talking friends?". This question can be simply reduced to calculating the similarities in a k-hop network. A k-hop network connects all those user pairs which can be reached from each other using k or less number of edges (hops) in the original network. Given a k-hop network, the $(k+1)$-hop network can be constructed by doing matrix multiplication of the adjacency matrices for the k-hop network and the original network M. A k-hop network in general would be much bigger than the original network and one needs to work on a randomly sampled subset of the network, where sampling is uniform over all the user pairs connected in the network. Getting an unbiased sample efficiently and analyzing how similarities die down as k increases is a part of the future work. Note that for a connected network, as k approaches N (N being the total number of users in the network) the similarity values will reach the random pair similarities.

4.6 Summary

Summarizing the results, we showed that people who talk to each other on the messenger network are more likely to be similar than a random pair of users, where similarity is measured in terms of matching on attributes such as queries issued, query categories, age, zip and gender. Further, this similarity increases with increasing talk time. The similarities tend to decrease with increasing average time spent per message. Also, we showed that even within the same demographics, people who talk to each other are more likely to be similar. Finally, as we hop away in the messenger network, the similarity still exists, though it is reduced.

5. RELATED WORK

Understanding the relation between the nodes and edges in a social network is an active topic of research in the areas of sociology and social networking, as well as in computer science.

In social networking, the idea that people with similar characteristics tend to be connected is called homophily. McPherson et al. [15] give an excellent review of work done on homophily in real-world networks. In their paper, McPherson et al. argue that additional research on homophily should be done, particularly with regard to how it affects the evolution of the social networks over time. We hope that our work can be a good starting point for understanding homophily on the Internet, and plan to look at the time-evolution of the network in future work. Sproull and Patterson [19] discuss how the participation in online communities might affect the every day lives and behavior of the people in the physical world. Our work can be seen as an experimental approach to analyzing these effects in the context of demographics and personal behavior (keyword searches) of the involved users.

There are a wide variety of real-world social networks that have been studied extensively in the literature. For example, networks involving sexual relations and disease [2][8]. Typically, though,

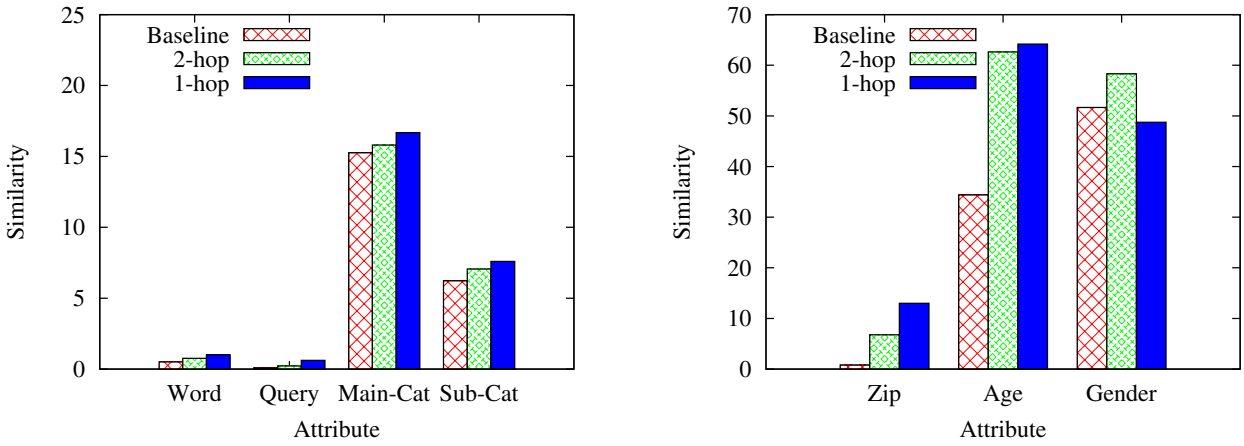

Figure 10: Similarities(%) in a 2-hop network: query attributes(left) and personal attributes(right)

they are fairly small, consisting of tens to hundreds of nodes. Of the research relating the characteristics of people with their social connections, we are not aware of any that was done on a real-world network of the size undertaken in this paper.

In the realm of statistics, a variety of work on modeling social networks also takes advantage of these principles. For example, Handcock and Raftery's [11] model for social networks incorporates assumptions about transitivity in link structure (if A is connected to B and B is connected to C, then A is likely connected to C), and attribute homophily (if A and B have similar attributes, they are likely to be connected). Given the model, inference on a partial network can be done to estimate unobserved links or attribute information. Work on modeling social networks has also been applied to viral marketing, with a series of recent papers that attempt to mine how important each node is in propagating certain ideas or innovations through the network ([6], [18], [12],etc.) as well as understanding the dynamics of a viral marketing system [13].

On the Web, it is often assumed that pages that are connected to each other are likely to be about the same topic. This "Web homophily" can be used to advantage in, for example, finding communities of Web pages [9][10] and the ranking of Web pages [17]. In the former, the homophily means an algorithm can find clusters of Web pages that are on a similar topic by looking only at the links structure. In the latter, it is assumed that a link between a pair of pages that are on the same topic is a "stronger" link, and should therefore be more highly regarded in the PageRank computation.

Because we have found homophily in the instant messaging network, it is interesting to apply the techniques used on the Web to the social network setting. The work on Web communities suggests that we may be able to find clusters of people with similar interests simply by looking at the structure of the social network. The work on ranking implies that we might want to consider social connections between users with the same interests to be "stronger" when determining, for instance, what people have high network influence, or are central to the network. We hope to study these effects in future work.

Leskovec and Horvitz [14] also performed an analysis on messenger data which overlaps with the data used in our experiments. While some of the demographic analysis is similar, our analysis focused primarily on measuring the similarity of users based on their interests, expressed by their search queries. We see their work as complementary to ours.

6. FUTURE WORK

There are many directions for future work in this area. We would like to experiment to see whether the positive correlations between keyword search similarity and IM talk time extend to the case of whether users click on advertisements as well. The hypothesis is that if one user clicks on an advertisement, and is connected to another user in the IM network, then the other user is more likely to click on the same ad. More generally, we hope to build a predictive model for both what searches the user is likely to make, as well as what advertisements the user is likely to click on, given who he/she talks to and the characteristics of those users (what they search for and what ads they click on). This could be used to personalize search engine results, or even suggest novel queries to the user that they may not have thought of themselves.

In this paper, we considered only chat sessions of two users. We would also like to experiment on multi-user chat sessions, to see if the correlations found in this paper exist (or might even be stronger) in such situations. It may also be possible to use the shared information about two users who are currently chatting in order to suggest a third person that they might be interested to join their conversation. We hope to build a model for when two users invite a third person into their chat, to see whether we can predict who that person would be given the interests and demographics of the two users already chatting.

Because the connections in a social network are based on communication, information tends to flow through them. We hope to study the query behavior of the users through time, to discover what types of queries (for instance, sensational news) tend to spread through the network, and what other queries (for instance, medical) do not. Such a model would capture the topological, as well as chronological properties of this spread, and identify the key users that influence this process.

There is some work on classifying queries based on their temporal characteristics (such as time of the day, day of the week etc.) [4] and type based on if it is a navigational, transactional or informational query [3]. We hope to incorporate these effects into our model as part of the future work.

Finally, we would like to examine whether the correlations discovered here are found in other domains, such as online gaming environments and social networking sites (such as Facebook and MySpace). We expect to find that, as with instant messaging, users who are connected in these networks will also be similar to each other. It will be interesting to compare the relative magnitude of

similarities across the various axes to identify how each domain differs.

7. CONCLUSION

In this paper, we showed that users who talk to each other in an IM environment are significantly more likely to share interests than a random pair of users. Our analysis is based on a probabilistic model over users and their attributes and relations. The similarity between users strengthens with the amount of time they spend talking to each other, and also holds for users who have a friend in common but do not necessarily chat with each other. To the best of our knowledge, this is the first experimental study of its kind. Though the results presented in this paper are preliminary, we believe that they demonstrate significant promise for further research in this area, paving the way for many advances in existing and novel applications for the Internet.

8. ACKNOWLEDGMENTS

We are thankful to Jure Leskovec and Eric Horvitz for the MSN Messenger network data, and useful discussions.

9. REFERENCES

[1] A. Abbott, editor. *The American Journal of Sociology*, 2006.

[2] P. Bearman, J. Moody, and K. Stovel. Chains of affection: the structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110:44–91, 2004.

[3] A. Broder. A taxonomy of web search. In *25th Intl. SIGIR*, pages 3–10, 2002.

[4] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In *14th Intl. WWW*, pages 2–11, 2005.

[5] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[6] P. Domingos and M. Richardson. Mining the network value of customers. In *7th Intl. SIGKDD*, pages 57–66, 2001.

[7] P. Doreian and T. Snijders, editors. *Social Networks*, 2006.

[8] K. Eames and M. Keeling. Monogamous networks and the spread of sexually transmitted diseases. *Math. Biosci.*, 189:115–130, 2004.

[9] G. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000.

[10] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[11] M. Handcock and A. Raftery. Model-based clustering for social networks. *J.R. Statist. Soc.*, 170:1–22, 2007.

[12] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence in a social network. In *9th Intl. SIGKDD*, pages 137–146, 2003.

[13] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. In *ACM Conference on Electronic Commerce*, pages 228–237, 2006.

[14] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *17th Intl. WWW*, 2008. To appear.

[15] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[16] L. Rainie and J. Shermak. Search engine use November 2005. Technical report, Pew Internet and American Life Project, 2005.

[17] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in PageRank. In *Advances in Neural Information Processing Systems 14*, pages 1441–1448, 2002.

[18] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *8th Intl. SIGKDD*, pages 61–70, 2002.

[19] L. Sproull and J. Patterson. Making information cities livable. *Communications of the ACM: Special Issue on Information Cities*, 47:2:33–37, 2004.

[20] S. WasserMan and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.