Extracting Social Networks from Instant Messaging Populations

John Resig, Santosh Dawara, Christopher M. Homan, and Ankur Teredesai
Center for Discovery Informatics, Laboratory for Applied Computing, Rochester Institute of Technology
Rochester, NY 14623, USA
jer5513,sgd9494,cmh,amt@cs.rit.edu

ABSTRACT

In the analysis of large-scale social networks, a central problem is how to discover how members of the network to be analyzed are related. Instant messaging (IM) is a popular and relatively new form of social interaction. In this paper we study IM communities as social networks. An obvious barrier to such a study is that there is no *de facto* measure for how closely any pair of members of such a community are associated to describe the link information. We introduce several such measures in this paper. These proposed measures are obtained solely from the status logs of IM users. The status log of an IM user is a list of pairs of the form (*time*, *state*), where *state* is an element of a small set, such as $\{online, offline, busy, away\}$, and *time* is the time at which the member switched into that state. Resig et al. show [12] that, in spite of their simplicity, status logs contain a great deal of structure. Since any pair of IM users can instant message each other only if they are both online at the same time, it seems reasonable to guess that any two IM users that are frequently online at the same time may in fact be frequently instant messaging each other. This hypothesis forms the basis of each of our association measures. For a chosen population of IM users, we compare the social networks obtained using our relationship measures to the social network formed in LiveJournal (www.livejournal.com) by the same population. LiveJournal is a blogging community that allows users to explicitly name other LiveJournal users as associates. The network obtained by these association lists thus acts as a control of sorts for validating our IM-based association measure.

1. INTRODUCTION

Instant messaging (IM) is a popular form of computer-based communication. By definition, IM is a communications service that enables its users to create a kind of private chat room with another individual that allows communication in real time over the Internet, similar to a telephone conversation but (typically) using text rather than voice. The instant messaging system alerts its users whenever somebody on their private list is online. Users can then initiate a chat session with that particular individual [1]. IM technology lets users communicate across networks, in remote areas, and in a highly pervasive and ubiquitous manner. Industrial and governmental organizations are very interested in understanding the nature of broad knowledge-sharing networks that exist within their organizations. IM communication is fast becoming a standard platform for such networks. Apart from a fundamental interest in knowing "who IM's whom and how often?" it is also useful as a test bed from a social network analysis viewpoint. From a data mining perspective, IM produces data at many levels of detail, ranging from state-change logs to text messages, and the data at each of these levels are rich in information. The problem of collecting, analyzing, and exploring this data has, until recently, gone mostly unexplored. Even the right questions to ask of them are not yet established, to say nothing of the algorithms required to efficiently answer the questions once they are posed. The IMSCAN framework is one such attempt to formulate and attempt solutions for such questions.

In this paper we focus on the particular problem of how to extract and analyze social relationships between the users of an IM service using the IMSCAN framework. A collection of such relationships between members of a population is called a *social network*. Social networks are widely studied, although often they are notoriously hard to analyze in any great depth. There are innumerable ways in which overlapping social networks can be derived from a population. This derivative is primarily dependent on the metrics used to determine the relationships. For instance, in a given group of people, any two people A and B could be considered related if A is the parent of B or if A knows B on a first-name basis or if A and B ever during June 2004 dined in the same restaurant. Relations can be either bi- or uni- directional. They can also be weighted; For instance, we could declare that the degree to which A and B are related is the number of times during June 2004 that A and B dined in the same restaurant at the same time. In each case, however, when we talk of a social network, we usually intend for the relation defining the network to indicate the degree to which some kind of meaningful social relationship exists between the members of the network (in the case of non-weighted relations the degree to which two members are related is either absolute or nonexistent).

LinkKDD'04, August 22, 2004, Seattle, Washington, USA.

A major problem one encounters in considering social networks among IM users is the lack of a *de facto* standard for what constitutes a relationship between any pair of users. Some IM services, such as AOL, provide client software that allows users to designate lists of ”buddies.” In this setting, it would be natural to say that two users are related if one lists the other as a buddy. Unfortunately, buddy lists are not published, so in order to obtain a collection of them one has to contact each author of each list. Thus, it is not practical to use buddy lists as the basis of a large-scale, IM-based social network.

Fortunately, many IM services, including AOL, constantly track the state of each user relative to the IM service. At any given time, each AOL IM client declares itself to be in one of four states: online, busy, away, or offline. This status data, along with the time at which a given client transitions from one state to another, is published electronically and is available online to any user of AOL’s IM service. It is thus possible to track the state of a population of AOL IM users over a period of time. In spite of the simplicity of this data, Resig et al. [12] showed that a great deal of structure exists in a typical population of AOL IM users that are monitored over a period of time.

It seems reasonable to us to use these status logs as the basis of a measure of the degree to which any two AOL IM users are related. For instance, any pair of such users can instant message each other only if they are online at the same time. In this paper, we propose several measures of the degree to which IM users are related based on how frequently they are online at the same time.

Before using our measures to extract social networks, whose structure would then be further analyzed, it is important to know just how well they represent any kind of real social bond between users. For instance, just because two users are frequently online at the same does not mean that they are ever talking to each other, much less that they know each other. We assume that users can communicate with each other only if they are in the online status. Some IM services allow users to be in ”Away” status or ”Busy” status and communicate. This does not preclude the methods we describe, though we do not consider it in our current analysis. We attempted to validate our measures by comparing them to another social network, which comes from the blogging web site LiveJournal over the same set of users whose IM usage we study. LiveJournal lets users publish buddy lists online. Thus it provides us with a social network where the relationships are stated explicitly by its members, and thus seems like a reasonable source of control data. In this paper, we present our measures and compare the relationships extracted by them to the relationships that exist in the LiveJournal network.

Organization: In the next section we describe the data collected and outline some preliminary terminology. We then describe two experiments we conducted to generate the community graph based on IM status log correlations. In Section 4 we describe a clustering based approach to link discovery and compare the links discovered with the Livejournal.com dataset. We conclude with a section on the results obtained and discuss the challenges for link discovery in IM networks.

1.1 IM Traffic Mining: Privacy Concerns and Cyber-Security Issues

Privacy concerns and cyber-security are two closely related issues. Governmental and intergovernmental organizations such as the G8, the US Government, and the Council of Europe (CoE) have been working to ensure lawful access to publicly available traffic data; yet all three have been criticized for adopting ambiguous and problematic policies and closed door approaches [13, 4].

Within all existing communication media there are two broadly defined types of monitoring systems: 1) traffic-based and 2) content monitoring. Telephone call monitoring and analysis of call data has been under the purview of government organizations for a long time, and in most cases traffic monitoring is considered less invasive. The directives within the recently tabled U.S. Government’s Patriot Act intend to bring cable and Internet service traffic under the same purview as telephone networks. Its a major concern if such surveillance attempts to eavesdrop on the actual content being communicated. The CoE ”‘Convention on Cyber Crime”’ defines traffic data as “any computer data relating to a communication by means of a computer system, generated by a computer system that formed a part in the chain of communication, indicating the communication’s origin, destination, route, time, date, size, duration, or type of underlying service.” [5] The IMSCAN framework is not invasive and does not have content monitoring capabilities. There are several issues in IM mining that are pertinent for link discovery and graph based social network analysis. Ability to understand and mine instant messaging networks is an important problem from the cyber-security perspective, and displays several challenges including but not limited to: graphical representation, relationship extraction, link-discovery-based community detection, and pattern learning [11]. Graphical representations of overall IM networks require techniques for processing large graphs. Relationship extraction based on sparse conversational evidence requires estimating who talks to whom based on probabilistic models.

2. DATA OVERVIEW

Instant messaging, a social-based communication medium, provides us with extensive social network information concerning its users. Concerning Instant Messaging social network data, many different media exist from which pertinent information can be extracted.

There exist two major types of link data, as related to instant messaging networks. The most useful, and hardest to acquire, of which are buddy lists. Whenever a user signs up with an instant messaging network, they are given a list in which they can mark users as being a friend. This information is, generally, stored on the central messaging server and is distilled to the client, once the user has authenticated themselves. Until that point, the link information is kept completely private. Another form of link information, that is more readily available, is that of 3rd party social-networking web sites, which allow users to provide their instant messaging contact information in their personal profiles. The most interesting aspect of the 3rd party web site is the fact that

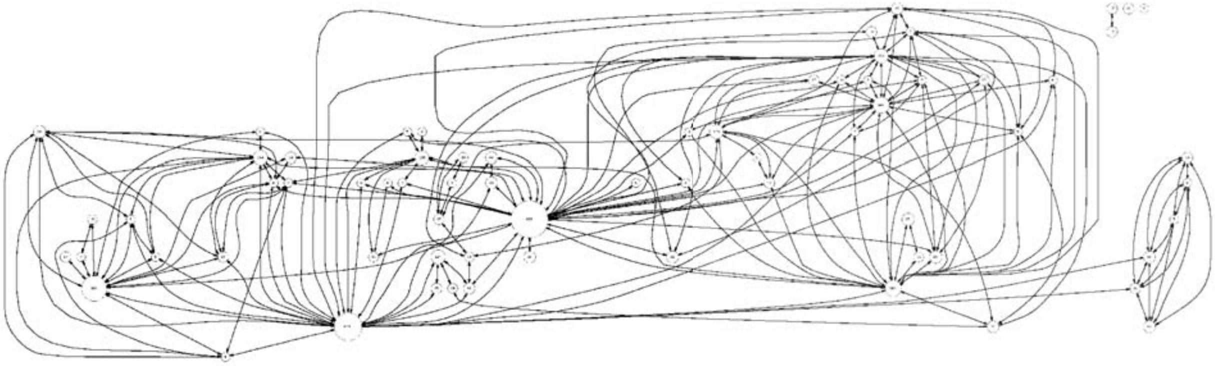

Figure 1: Graph of communities within the pruned 15-in and 15-out link social network. Communities are represented as circles, the number inside of them representing the number of users within the community. The links passing from the circles represents the flow of loose connections (only an in or an out link) between the communities.

Sample Tracking Data:

Time (s)	User	Status
15	User2	Online
45	User1	Offline
60	User1	Online
130	User4	Away
160	User5	Idle

the information provided is readily, and easily, accessible to the public.

We collected data from a 3rd party social network (due to its easy accessibility) called LiveJournal. LiveJournal users have the ability to mark other LiveJournal users as being a 'friend' of theirs, thusly creating a social network of associated users. This associations form the basis for links within the social network, with each of the users being a vertex. Users can also provide information pertaining to their personal account, such as their Instant Messenger user names. Over 200,000 user names and their associated IM names were collected.

Sample Social Network Data:

User	Friends
User1	User3 User5 User7 User18
User2	User3 User4 User8
User3	User1 User2 User19 User30 User31
User4	User2 User6 User19 User20

In order to prune the dataset, and find an ideal group of users to collect status data on, users with a minimum in and out link degree of 15 were chosen, leaving us with a group of 4878 users. A group of users, of this particular size, was ideal due to the bandwidth constraints that currently exist within the IM tracking network. In order to keep the amount of traffic to a manageable level, the number of selected users was kept below 5000. Once these users were selected, they were then tracked using our IM tracking framework [12] for 25 days.

Pruned Network Information:

Vertices	4878
Edges	309953
Diameter	22
Avg. Shortest Distance	6.009

One method which was used to analyze, and view, the distribution of users within the LiveJournal social network was the development of 'communities', Figure 1. Communities are defined as: Groups of users who have both an in and out link to each other. Unlike cliques, all users of a community do not have to have reciprocated links with every other user, simply having reciprocated links with an user within the community is enough to warrant being a member of the community.

3. CORRELATION BASED LINK DISCOVERY

3.1 Overview

In this section, we attempt to correlate user activity in AOL instant messaging status logs with the distribution of links in the LiveJournal graph. One motivation for studying this problem is that, if instant messaging user activity did correlate in some way with the links in the LiveJournal graph, we could use the LiveJournal graph to train a system that extracts links between users using only the instant messaging status log as test input.

We describe two experiments: In the first, for each user in the LiveJournal graph, we compare the number of seconds that user was online on AOL IM to the out-degree of the node in the LiveJournal graph corresponding to that user.

In the second, for each pair of users in the LiveJournal graph (not necessarily linked), we measure the degree to which the pair's IM online is synchronized to whether the pair is linked in the LiveJornal graph.

3.2 Experiment 1: Comparing IM online presence to LiveJournal popularity

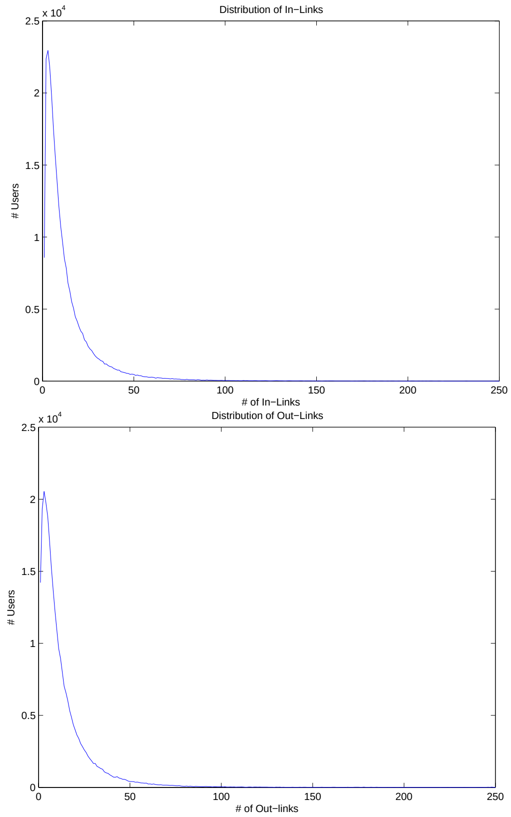

Figure 2: In and Out-Degree distributions for the LiveJournal social network data.

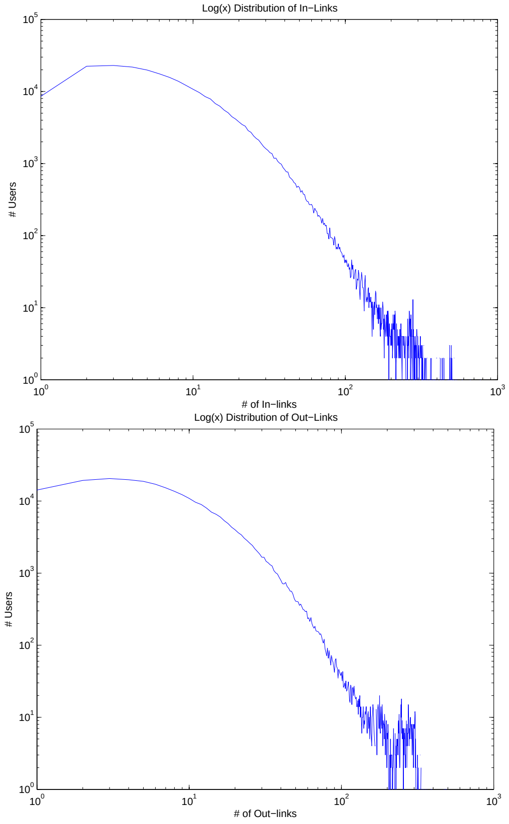

Figure 3: Log(x) In and Out-Degree distributions for the LiveJournal social network data.

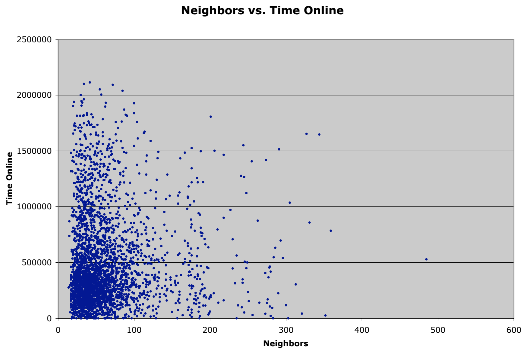

Figure 4: The amount of time the user spent online versus the number of out-links the user has in the LiveJournal graph. There appears to be no correlation between the amount of time spent online and the number of out-links.

In the first experiment, for each user, we count in the instant messaging activity log the amount of time in seconds the user was online and the number of times the user changed his or her status to online from some other state. We matched this data to the number of out-links in the LiveJournal graph the user has. We define the users that these links point as the *neighbors* of that user. Figures 3.2 and 3.2 graph the amount of time the user spent online (respectively, number of times the user moves into the online state from some other state) by the number of out-links the user has in the LiveJournal graph. The graphs show that there appears to be no correlation between the amount of time spent online and the number of out-links.

3.3 Experiment 2: Comparing the degrees to which pairs of IM users are online at the same time to LiveJournal linkage

In the second experiment, we measure the degree to which each pair of users is, according to the IM status log, online at the same time. When we say that the degree to which a pair of users (x, y) is online at the same time is n, we mean that there are n points in time at which x or y goes (or both x and y go) online from some other state, and both x and y are online. We use this measure of frequency rather than simply counting the number of seconds x and y are online at the same time because if x and y are both online at time t and the time-sampling frequency is small then it is very likely that x and y will still be online at time $t + 1$.

After measuring the degree to which each pair of users is online at the same time, we divide the pairs into two subpopulations: one of all pairs that are linked in the LiveJournal graph, and one that has all pairs that are unlinked. We then compare the distributions corresponding to each population of the degrees that pairs of users are online at the same time. Figure 3.3 shows that the two distributions are essentially the same. Note that, because the populations of the two groups is very different (there are approximately 93 times as many pairs of users in the unlinked subpopulation as there are in the linked subpopulation). Although the tails of the two populations are different, we suspect this is because the unlinked population is much larger than the linked population and so has a smoother distribution.

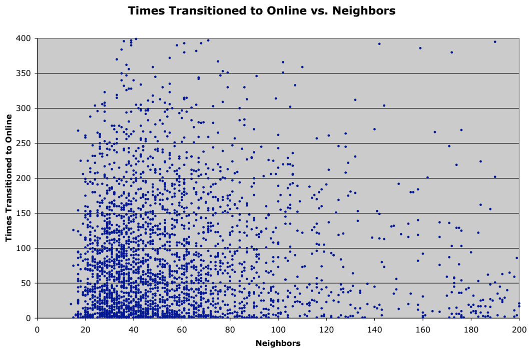

Figure 5: The number of times the user moves into the online state from some other state versus the number of out-links the user has in the LiveJournal graph. There appears to me no correlation between the number of times the user moves into the online state from some other state and the number of out-links.

4. CLUSTERING BASED LINK DISCOVERY

4.1 Overview

This section is devoted to the study of techniques to recovering associations between users from their observed behavior by first clustering [10] them based on the status logs.

Clustering allows the reconstruction of an optimal set of groups to represent the original dataset. The groups are formed such that members within each group are as similar to each other as possible and members in different groups are different from each other. The degree of similarity, or the metric is based on some characteristic of the dataset.

We examine the argument that the links between users can be reconstructed using clustering. The distance metric is reasonably based on similarities in the behavior of any two users. The more inter-linked the behavior of two users, the greater the possibility that they are related in a larger frame of reference.

The objective is to cluster IM users based on their behavioral patterns. Similarity is assumed to exist between every pair of members in the same cluster.

4.2 Formulation

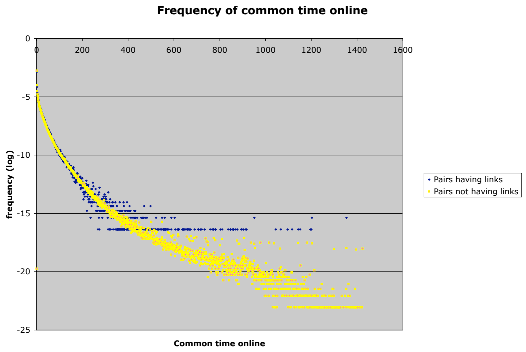

Figure 6: This graph plots two subpopulations. The first contains all pairs of users who are linked in the LiveJournal graph. The second contains are pairs of users who are not linked in the LiveJournal graph. The x-axis represents the amount of time a pair is commonly online via the "degree of time online" measure described in the text. The y-axis represents the frequency (as a fraction of the size of the subpopulation the pair is a member of) that a pair having the given x value exists. The graph shows that the distributions for both populations are not significantly different.

Consider a data set X consisting of N users $(a_1, a_2, a_3, ..., a_N)$ under observation over a period of time (range $[0, T]$). Each tuple represents user-status at time t.

The objective is to group users a_1 to a_N into clusters based on similarity in status-change behavior. Users within the same cluster suggest a strong association and are likely to have a link between them. In other words, clustering would provide us with a single or multiple valued function $J(a)$ such that, $J(a_i)$ assigns user a_i to some clusters out of $C_{1..k}$ clusters, where k is the number of possible clusters. Such a clustering scheme is referred to as soft clustering since user a_i is allowed to be a member of more than one cluster. For the purpose of link discovery from clusters, we define a link to exist between a pair of IM users as follows: Definition Users a_i and a_j are considered to be linked if, $\exists$ some cluster C_k, such that, both $a_i \& a_j \in C_k$. The above definition suggests the following requirements of a clustering algorithm to be applicable for link discovery:

1. Since the notion of links is symmetrical the distance metric used in the clustering has to out put a distance between two users.
2. The method should be capable of providing a global perspective when determining the relationship between any two users. In other words, the method should be able to leverage the transitive nature of relationships between attributes (here users) as suggested by previous efforts in clustering categorical data [2, 6, 7, 14].
3. The degree of similarity is the least when the behavior of the two users being compared, is unrelated. The degree increases in proportion to an increase in the correlation

4.3 Information-theoretic co-clustering

Information-theoretic co-clustering [3] is an implementation of the co-clustering class of algorithms [8]. It is a non-hierarchical and iterative method, similar to the k-means clustering algorithm [9]. Information-theoretic co-clustering algorithms simultaneously cluster both dimensions of the data set. The algorithm provides a clustering function that clusters rows, incorporating column cluster information and vice versa.

Information theoretic co-clustering was the comparative technique of choice to meet some of the requirements imposed for determining similarity between users. It satisfies the requirements for both an internal and external method. The internal method provides a view to compare how similar two users are provided the actions are unbiased and independent and the external method can be used to compute the similarity under the dependence constraints as mentioned earlier. However, the drawback is that it can only provide hard or non-overlapping clusters. Like k-means, the algorithm also requires to be seeded with an initial value describing the number of groups for each dimension k_{row} and k_{col}. In that respect, it is a sub-optimal solution to our problem[1].

4.4 Experiments and results

The user status log is transformed to highlight interconnected behavior between users. This information is used for the further experiments.

The strategy adopted to discover links is to find patterns in the manner in which users change to the online state and stay in that state for an extended period of time. Based on our knowledge of the instant messaging protocol, a user is only likely to be interacting with another user if they are both in the online state. The assumption that continued interaction may only take place when both users are online is for convenience, and does not accurately reflect the usage of the IM network.

Data set X is pre-processed to give an intermediate relation $\hat{X}_l$. The relation has dimensions N by N. Each cell x_{ij}, in the relation is a measure of the degree to which the behavior of users a_i and a_j correlate.

One way to represent the behavior of every pair of users is to define x_{ij} as the number of times users a_i and a_j were found to be online together, every t seconds over the entire time period $[0, T]$ spanning the experiment.

Optionally, this score could be weighted by:

1. Reducing it in proportion to the amount of activity in the background. Thus, scores accumulated will be a lot lower if two users are online during peak periods of activity.

[1] Determining the utility of other clustering techniques is an area for further exploration.

k (rows)	k (columns)	links recovered	links verified	precision (%)
50	50	258,616	2,890	1.117
100	100	169,292	1,926	1.113
250	250	94,116	1,179	1.252
500	500	49,720	723	1.454
750	750	31,080	441	1.418
1000	1000	25,458	335	1.315
1500	1500	14,848	235	1.582

Table 1: Clustering based similarity is dependent on the number of clusters.

2. Reducing it if the two users stay online together for increasing amounts of time. Thus, an event where two users come online together for a short period of time has a greater weighting over a pair that remain online together for relatively greater lengths of time.

The score is intended to be a function of the probability of a pair of users interacting with each other.

We then proceed with clustering data set $\hat{X}_l$ to give groups of users for various values of k. It can be observed from the results described in Table 4.4 that as the size of k increases the strength of similarity between users declines. This in turn validates the results we obtained in the previous section since it clearly indicates the transient nature of the definition of similarity in this context. Hence, under the assumption of independence of user actions, if direct or internal distances between pairs of users would have been good measures, the memberships would have remained invariant as k increased. Since the notion of similarity is not correctly captured by the existing clustering methods, we argue that more effort is required to correctly define the notion of similarity (perhaps focusing more on external distances as suggested by Das et al. [2] and Ganti et al. [6]) for categorical datasets such as IM status logs.

4.5 Discussion

LiveJournal provides it's own plane of reference within which to create and grow social communities. However, the primary motivation is not the definition of general communities, but communities of bloggers and blog readers. Therefore, LiveJournal is not a definitive source for verifying links and it is difficult to draw conclusions. Nevertheless, the very poor precision of the proposed strategies in applying clustering is disappointing. Note that it is not possible for two users to communicate with each other, if they are not in the online state together. This of course only implies that if two users are online together, the probability that they are linked increases. We conclude that further efforts are required to accurately model this relationship.

Additionally, investigation is also required into clustering (including soft clustering) methods before any claim of success can be made in the recovery of information of social communities using clustering. It is also important to provide the ability to determine the accuracy of any proposed model that is used to discover patterns in the IM data for link recovery. The authors are actively researching alternative sources that can help recover and verify links.

5. CONCLUSIONS

In this paper, we defined and investigated two metrics to formulate the process of link discovery in social networks such as instant messaging services. We first introduced the problem of defining the various ways in which relationships can exist between a set of instant messaging users. Then we succinctly described the broad policy on traffic versus content monitoring and placed the issue under a neutral perspective. We highlighted the fact that posing important questions and providing algorithmic solutions to them is in itself a very interesting and challenging task for this domain. The experiments and results for the two metrics demonstrate the utility of studying these issues to gather a better understanding of the algorithmic issues. There is considerable interest and scope for future work on the subject of instant message mining, including the use of this dataset to formulate problems in determining communities of users, extracting relationship information more effectively and visualization of evolving networks to cite a few possible directions.

6. REFERENCES

[1] Definition of instant messaging: http://www.webopedia.com/.

[2] G. Das, H. Mannila, and P. Ronkainen. Similarity of attributes by external probes. In *Knowledge Discovery and Data Mining*, pages 23–29, 1998.

[3] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98. ACM Press, 2003.

[4] A. Escudero. Contribution to the european union forum on cybercrime. Technical report, Council of Europe, Brussels, Nov. 2001.

[5] A. Escudero-Pascual and I. Hosein. Questioning lawful access to traffic data. *Communications of the ACM*, 47(3):77–82, March 2004.

[6] V. Ganti, J. Gehrke, and R. Ramakrishnan. "cactusclustering categorical data using summaries". In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 73–83. ACM Press, 1999.

[7] D. Gibson, J. M. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamical systems. In A. Gupta, O. Shmueli, and J. Widom, editors, *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, pages 311–322. Morgan Kaufmann, 1998.

[8] J. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, March 1972.

[9] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[10] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[11] R. Popp, T. Armour, T. Senator, and K. Numrych. Countering terrorism through information technology. *Communications of the ACM*, 47(3):36–43, March 2004.

[12] J. Resig and A. Teredesai. A framework for mining instant messaging services. In *Proceedings of the 2004 SIAM Workshop on Link Analysis, Counter-terrorism, and Privacy*, Lake Buena Vista, Florida, April 24 2004.

[13] U.S.Delegation. Discussion paper for data preservation workshop. In *G8 Conference on High-Tech Crime*, May 22-24 2001. Tokyo.

[14] Y. Zhang, A. Fu, C. Cai, and P. Heng. Clustering categorical data. In *In Proceedings of the ICDE*, page 305, 2000.