Instance Based Social Network Representation

Yoav Artzi
Computer Science & Engineering
University of Washington
yoav@cs.washington.edu

Amit Levy
Computer Science & Engineering
University of Washington
levya@cs.washington.edu

ABSTRACT
With the prevalence of personal information sharing on the Web, users are becoming more concerned with their privacy online. Social networks, in particular, represent a new challenge in this regard since they combine the sharing of highly personal data with ease of access. While most of the current discussion addresses the problem of providing a simple interface to configure complex privacy settings, this paper focuses on the user's ability to reason about how privacy policies affect their data. We show that it is possible to concisely represent a large subset of a social network using only a few representing members. Moreover, we argue that such a representation can inform users about the implications of privacy policies on who can see their data.

Author Keywords
social networks, clustering, online privacy, community detection

INTRODUCTION
In the past few years there has been an explosion in the social Web with the rise of social networks and content. An obvious result of this development is the increase in personal information migrating online, rendering privacy a critical issue. This trend, in combination with the ubiquity of search engines, has made private information easily accessible, further increasing the urgency of addressing privacy issues. Recently, concerns over users' control of their privacy in social networks like Facebook [3] have gained significant media attention [1] [11] [15]. Most of the discussion has focused on abrupt changes to privacy policies and complicated configuration interfaces [4] that confuse and discourage the user from restricting access to their data.

Another important aspect of privacy in this setting is the ability of users to reason about how their privacy settings manifest with respect to specific resources. For example, who has access to a particular picture, video or post? Empowering users to answer this question easily might have an affect on what they decide to share with their social network, as well as their demand for stricter or more open privacy settings.

This paper aims to present a method of communicating information about the state of the user's social network. We propose a general purpose approach for representing social networks in a concise manner. Our goal is to leverage the user's knowledge of the network to simplify and minimize our representation, while retaining as much information as possible. Such presentation of the network can illustrate exposure and privacy settings to the user in a compact and accessible manner.

MOTIVATION
Displaying the set of representing members can help users develop intuition about the larger subset of their social network being represented. As opposed to displaying the entire network (using a list, for example), instance-based representation consumes very little screen real estate and requires significantly less attention from users. Therefore, users are less likely to ignore such a lightweight representation and more likely to extract meaning from it.

This method can help illustrate the exposure of users' resources. In the case of photos, for instance, we wish to express the subset of the user's social network that has access to a particular photo. Representing members are displayed along the photo to indicate what part of the social network has access. Such a system can be combined with a drill down interface that aids in further exploring the represented clusters.

To illustrate the practical potential of this method, let's examine the case of Alice. A few weeks ago Alice called in sick from work, while she actually traveled to Paris. Upon return, she wants to share her experiences with friends by posting her Eifel Tower photos. Since many of her colleagues are Facebook friends of hers, including her boss, they could see the photos. However, while posting the new album, a small list of friends appear along side the photos indicating who can see the album. On this short list, Alice notices her co-worker Bob. She quickly realizes that people from work can see the album, including maybe her boss. Knowing this, Alice decides to be more cautious publishing her photos.

In general, this approach can aid in representing a large subset of the social network when screen real estate or user time is limited. Some immediate examples are (see figure 1 for screen grabs from Facebook):

- A large number of people that "like" an item.
- The mutual friends of two users.
- Many friends tagged in an album.

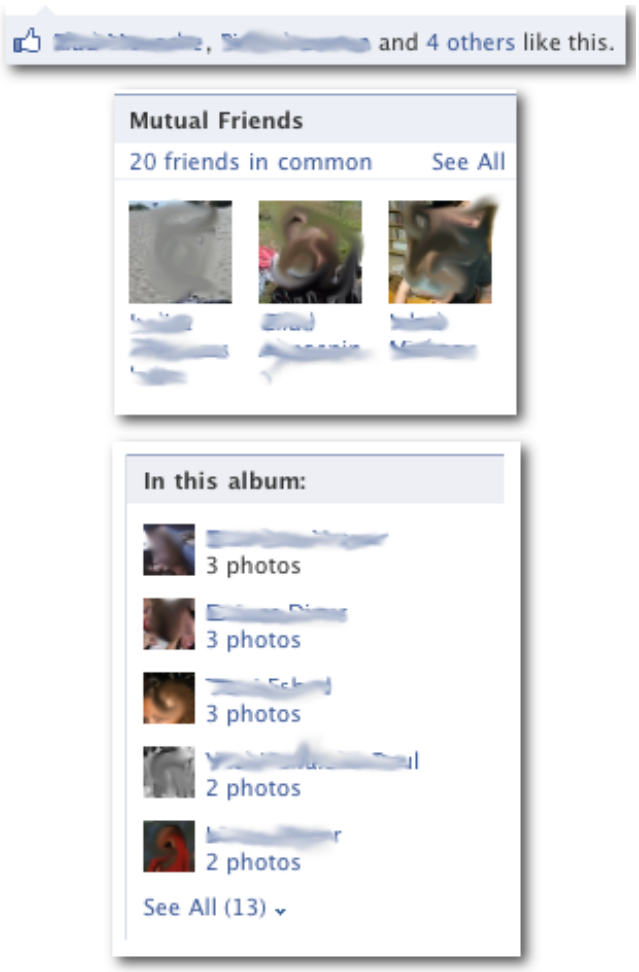


Figure 1. Whenever limited screen real estate is available, instance based social network representation can be used to create better coverage of the represented network.

RELATED WORK

Online Privacy

Palen and Dourish [12] identify privacy as an HCI problem. Their analysis approaches the issue through traditional analysis of privacy in everyday life. Their emphasis on the dynamic nature of privacy, highlights the need for constant awareness of privacy state and exposure. Although this comparison highlights the importance of good privacy management as well as some of the unique problems of the Web, it does not capture the scale in the context of the social Web.

Lipford et al. [10] acknowledge the problem created by users disclosing large amounts of personal information on social networks. They suggest allowing the user to define different views of their profile. Each view is restricted to a subset of members of the social network: one view appears in search results, another for friends and so on. Although this approach successfully illustrates the separation between various scopes, it does not address issues highlighted by Palen and Dourish. Specifically, it does not tackle the dangerous mixture of persistent data and a dynamically changing network. Finally, their approach assumes a social network with a small number of well-defined sets of friends, which are either disjoint or display a containment relation. This assumption does not generalize well to social networks like Facebook.

Baatarjav et al. [2] analyze the information shared across social networks. The risk posed by rouge users highlights the need for better visualization of information exposure. They also propose a semi-automated privacy management system for social networks. Their system makes use of probabilistic approach based on information revelation of users across the network to suggest a more appropriate privacy policy. While such an approach can prove useful for generating better default policies, the variability in costs for different stake holders also requires the ability for users to monitor their privacy through better visualization.

A more personal approach to privacy was suggested by Lieberman et al. [9] in the context of webmail. They show that people manage their privacy more efficiently when presented with photos of the participating sides of an email correspondence. Furthermore, they consider cases such as emails addressed to many recipients or to a mailing list, and propose a system which avoids information overload in these cases. In their work they recognize the importance of visual representation and the challenge of handling large networks. We borrow from their observations and apply them in new settings.

Community Detection

The field of community detection studies grouping patterns of humans in social networks. Porter et al. [14] conducted a thorough review of the field. Existing algorithms for community detection range from traditional clustering techniques to specific partitioning methods. Community detection methods treat the social network as a graph with the vertices representing members and edges the ties between them. Such graphs can be weighted and even directional.

Social Ties

The concept of social tie strength was introduced by Granovetter [6]. According to Granovetter the strength of a tie is a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and the reciprocal services characterizing the tie. Granovetter concluded that not all ties are different: some are stronger and some weaker.

Tie strength in online social networks was explored by Gilbert and Karahalios [5]. They used features of the Facebook social network to predict tie strength between members. They identify features related to the network, communication between users and various other factors. Examples include the number of common friends, "wall" words exchanged and co-occurrence in photos. Their results show that such features can predict tie strength with high accuracy.

Network Features
Friendship connection
Number of mutual friends
Personal network sparseness
Groups in common

Communication Features
Wall words exchanged
Inbox messages exchanged
Days since first communication

User Profile Features
Age difference
Distance between hometowns
Mutual affiliation (networks)
Education difference

Table 1. A sample of features of the Facebook social network that can be used to predict tie strength.

Centroid	$\arg\min_u d(u, C)$
Cluster Edge	$\arg\max_u d(u, C)$
Connectivity Heaviness	$\arg\max_u \sum_v \omega(u, v)$
Connectivity Lightness	$\arg\min_u \sum_v \omega(u, v)$

Table 2. Cluster representing selection methods. C is the centroid of the cluster and d is a distance function.

INSTANCE BASED SOCIAL NETWORK REPRESENTATION

We generate instance based network representation from the weighted social graph G_x, where nodes represent friends, and edges represent connections between friends:

$$G_x = (V, E, \omega)$$

Where x is the current user and:

$$V = \{v_m | m \in SocialNetwork \wedge m \neq x\}$$
$$E = \{(u_m, v_n) | u_m, v_n \in V \wedge \omega(u_m, v_n) \neq 0\}$$
$$\omega(u_m, v_n) = TieStrength(m, n)$$

We use weights to express a spectrum of tie strengths between friends, since two friends may be connected in several ways, and since each connection has a different semantic meaning. Following Gilbert and Karahalios [5], Table 1 lists a few features of the Facebook social network that can be used to calculate tie strength. Following the creation of the graph, it is partitioned using a clustering detection algorithm. For each of the resultant clusters a representing member is selected. Table 2 lists a few possible methods to select the representing member.

The above process can be executed for the entire network of a user or for a subset of it. In the case of the entire network, it's possible to define this network to be the direct friends of the user only, or include friends of friends as well. However, in some cases the base network is actually a subset of the social network of the user. For example: when representing users that can view a photo, the subset will be the users who actually can see the photo considering the active privacy policy.

EXPERIMENTAL SETUP

We leveraged the Facebook developer API [7] to develop an online experiment that measures the effectiveness of instance based representation in approximating users' mental model of their social networks.

When a participant logs in to our experiment using their Facebook account, we collect uniquely identifying information (namely their user id), as well as a snapshot of their social network - their friends, the mutual relationships between their friends, and the networks and location of each friend. We decided to limit the network features used in constructing to the network to simplify the experiment and to conduct it within Facebook's licensing terms.

Direct friendships are always assigned a weight of 1. Location and network connections are assigned varying weights on a per-experiment basis of either 0 (in which case we do not consider the feature), 0.2, or 0.3. We vary the values of these weights in such a way that produced data with both of the features excluded, both included, and each feature included individually. Note that we chose to implement a symmetric weight function. Once a graph is constructed, our platform applies the Walktrap [1] community detection algorithm [16] on the user's social graph to divide friends into k clusters. [2] For each resulting cluster, a representative is chosen by finding the friend with the greatest sum of weights within the cluster (intuitively, the friend that is most *connected* to other friends in the cluster).

Since our social graph excludes the current user, it can happen that certain friends will not be connected to any other friends. This situation is more likely as less features are used. For example, if only the friendship feature is used, friends the current user has no mutual friends with, will be isolated. This means the clustering algorithm has no information about such users for the partitioning process. We devised two approaches to deal with such situation:

1. Cluster all the isolated members together and pick their representing member randomly.
2. Create a singleton cluster for each of them with the representing member being the isolated user for each such cluster.

In our experiment we chose to use the first option to maintain control on the number of generated clusters. Naturally, as more features are used, the likelihood of a member being isolated decreases.

The participant is then presented with ten questions based on the clusters and representatives generated from their social graph. Each question presents the participant with some subset of their social network (namely pictures of some of their friends) and asks them to categorize them in a mean-

[1] Walktrap [16] is a community detection algorithm for weighted social graphs by Pons and Latapy [13].

[2] The value of k depends on a step function based on the number of friends a user has and ranges from 7 for a friend count less than 50 to 20 for a friend count greater than 200

ingful way. The format of the questions is chosen from the three formats described below. Furthermore, a set of parameters upon which to cluster the participant's social graph is chosen for use in all ten questions, but for each question different clusters are chosen at random. Thus, the data gathered from a participant at the end of the experiment represents a data point with one set of parameters for the weights of location and network, and one type of question. Participants are not revealed the parameters used for their experiment.

We decided to distill our problem to a set of questions to avoid the difficulty of measuring the representation quality on privacy problems. First, privacy issues, although they carry high cost, are relatively rare. Second, Facebook's API separates applications, making it impossible to inject code to areas such as the wall or photo albums without having the user actively install a local script or program on their computer. We hoped to simplify the experimental environment, so we can test only the quality of our representation. Through the different question types we covered different aspects of representation quality.

We proceed to describe the three questions participants saw during the experiment, see figure 2 for screenshots:

Representing Member Selection Question

The participant is presented with the representing members of four different clusters from their social graph, as well as one randomly chosen friend. The user is then asked to choose which of the representatives are most related to the randomly chosen friend, or *none* if none of the representatives are related. If the participant chooses the matching representative the answer is considered correct, and if the participant chooses any other representative it is considered incorrect.

This question measures the accuracy of the clustering algorithm in carving out clusters which are well represented. That is, the algorithm is specifically penazlied for ambiguous clusters, which force the user to ''break ties'' between two representatives.

User Relevancy Question

The participant is presented with a similar setup as the previous question. Instead of asking to specify that a particular representative is related to the randomly chosen friend, we ask the participant to choose *yes* if *one* of the representatives is related to the friend, and *no* if none of the representatives are related.

This question measures the power of a *set* of representatives to communicate some larger sub-set of the social graph. In contrast to the previous question, the user must only indicate if the friend is represented by *at least* one of displayed representatives, and not distinguish *which* one.

Group Construction Question

In the third question type, we show one representative to the participant and a series of nine randomly chosen friends. For each friend, the participant is asked to indicate (by marking a checkbox or leaving it unmarked) if the friend is related to the representative. For this question, we choose the most adversarial scoring approach to our hypothesis. Answers are evaluated *all or nothing* – that is, if the model's prediction is off by at least one friend it is considered wrong.

This question asks the user to reconstruct a portion of the cluster given a small subset of the social graph. This is really measuring how well the algorithm approximates how the participant would cluster their friends.

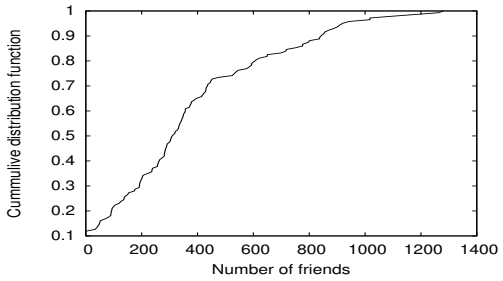


Figure 3. Network size. Our participants include Facebook users with a wide distribution of network sizes (ranging from 33 to 1280 friends). The median network size is 310 friends.

RESULTS

We collected data from 110 participants. A portion of the participants (31) did our experiment twice but with different settings. This gave us a total 140 data points. Participants were recruited online through Facebook and mailing lists, most were acquaintances of ours. Since the experiment was conducted online, participants logged in through a Web browser and answered the questions. Asking the participants to come to a lab was not likely to provide us more informations and would have significantly increased the effort required by them, probably decreasing our pool of participants. Presenting a relatively small number of questions (10) of a single type, also made the experiment simpler to complete and less time consuming. Participants received no compensation for their participation.

We define the prediction accuracy of a model for each participant to be the number of questions answered as expected, divided by the total number of questions. Across all participants, the mean accuracy was 0.79 with a standard deviation of 0.17. The median was slightly higher that the mean, at 0.81, indicating a slight skew in the data - this is expected for such a high mean since accuracy is distributed between $0 - 1$. Figure 4(b) shows the distribution of accuracy across participants.

The number of friends each participant had in their network varied greatly, and ranged from 33 to 1280. Figure 3 summarizes the distribution of network sizes. We account for this disparity by adjusting the number of clusters by which to divide a social graph to the size of the graph. We evaluated this method by analyzing the impact of network size on accuracy. As expected, there was no significant correlation.

We further evaluate the influence of network affiliation and

Figure 2. Question screens. From top to bottom: representing member selection question, user relevancy question and group construction question.

Question Type	Accuracy	Std Error
Member Selection	0.858	0.034
User Relevence	0.813	0.035
Group Construction	0.695	0.045

Table 3. Summary of least-squares regression for each question type.

geographic location of friends. Recall that both dimensions were used to increase the weight of edges between friends or introduce an edge where none existed. If two friends of a participant had the same network affiliation or the same location, the weight of the edge between those friends would increase by some amount. This amount was varied for each participant, and we collected a similar amount of data for each combination. The possible combinations were:

- Network weight: 0.0, location weight: 0.0
- Network weight: 0.0, location weight: 0.2
- Network weight: 0.2, location weight: 0.0
- Network weight: 0.3, location weight: 0.3

We could not find any statistically significant relationship between the network or location weights and the prediction accuracy of the model, when analyzing the whole data and when dividing it by question type.

We performed a mixed-model analysis of variance, treating all random variables (affiliation weight, location weight, network size and question type) as fixed effects, except the user ID, which was treated as a random effect. We removed affiliation weight, location weight and network size due to being insignificant. The omnibus test showed a significant main effect of the question type ($F(1, 140) = 13.96, p \approx .0003$). This prompted us to divide the data by question type and analyze each using least-squares fit. Still, location weight, affiliation weight and network size showed no significant interaction with accuracy. Figure 4(a) shows the accuracy for each question type and location/network weights combination. Figure 4(b) shows the distribution of accuracies for each question. Table 3 summarizes our analysis.

Limitations

We identify several limitation in our study:

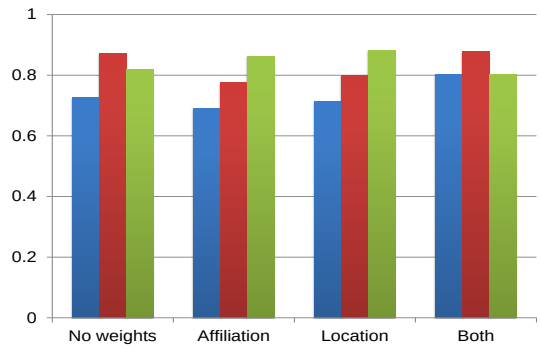
1. It was important for us to respect the privacy of our participants and obey Facebook's licensing agreement. Therefore we were limited to the official Facebook API. This restricted our access to data about our users and about the interaction between them.
2. We recruited participants from our circle of acquaintances due to the limited time available for data collection. Naturally, this created bias due to various properties kept relatively constant among our friends, such as age, education, socioeconomic background and geographical location.
3. We chose to conduct a between subject analysis, asking each participant. only one type of question with one set of weights for location and network. Part of the motivation for this choice was practical - we wanted the experiment to be simple and easy to complete. However, due to the large variability between users' social network properties, it is possible that a within subject study is more appropriate, and we consider this future work.
4. Despite the high number of potential participants (Facebook has over 400 million active users [8]), we collected a lower number of participants than we expected. This limited our ability to test all the possible dimensions of the problem and, naturally, harmed the validity of our results. One reason for this could be that people are becoming more concerned about their privacy and more reluctant to share their profile. This became apparent especially over the past few weeks with the inflation of Facebook privacy related news articles.
5. Due to several bugs we found in our original experiment platform, we were forced to dump some of the results. This, combined with the low turnout, forced us to re-use some of the users. However, recycling of users occurred across different question types, which were analyzed separately to minimize the damage to the results.

DISCUSSION

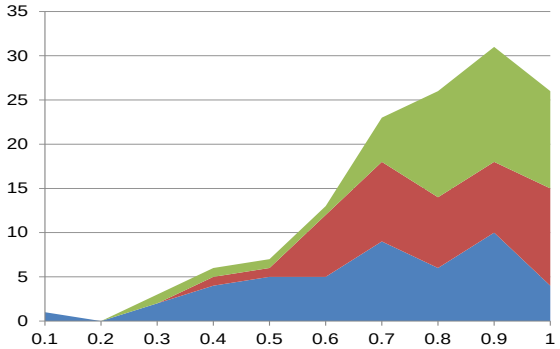
Our results show that instance based social network representation holds potential for concise representation of social networks. The goal of such representation is not to provide an accurate and complete image, but rather to provide users with intuition about the relevant subset of the social network. In this sense, the representation displayed to the user is an aid targeted at the network owner and not independent on its own. The high average success rate of participants across the various question types shows that our representation allows various kinds of inference that are required to comprehend the larger social network represented.

The different question types illustrate the inherent complexity of measuring the success of such intuition-directed methods. For the first two question types: representing member selection and user relevancy, we received similar results. The third, and more complex, question showed slightly lower mean success rate and higher standard deviation. This can be attributed to its higher complexity: the first two question included a single decision, whereas the third was composed of nine separate decisions. The higher success rate of the simple questions shows that our method is more successful at helping the user answer questions about specific users. For example, let's examine the case when the user doesn't want a specific friend to see a photo that might embarrass that friend. Seeing a representing member related to that friend along the photo, will increase the user awareness of the chance the photo is viewable by that user.

We were slightly surprised by the location and affiliation properties not influencing the quality of the representation. Since both these fields are not mandatory and some users choose to leave them empty. We believe that the approach of using extra properties of both the network and user information to construct the weighted social graph deserves further analysis.



(a) Mean accuracy of each question type separated by network/location weight combinations.



(b) Stacked histogram displaying distribution of accuracy by question type.

Figure 4. Accuracy by question type and weights. The green areas represent the *Representing Member Selection* question, the red areas the *User Relevance* question, and the blue areas the *Group Construction* question.

Properties that are created during normal social activity might be more helpful. Such properties include wall posts exchanged between users, private messages and the content of both. Unlike location and affiliation, these are generated during normal activity and the user is not required to actively supply them. Intuitively, it's easy to theorize how such properties form evidence for tie strength. For example, if two people write on each other's walls often, there is likely a strong connection between them. Unfortunately, such properties are not available through the official API and scraping them stands against Facebook's licensing agreement.

Network properties may also prove helpful. These include for example: the size of the networks of the members of a tie, the sparseness of their network and the number of their mutual friends. For example, two users who have 20 mutual friends are likely to have a stronger tie than two users with only two mutual friends.

The two observations above also undermine the assumption of ties being symmetrical. For example, if one user has five friends and another 500, it's possible that the tie between them carry much weight for the user with the smaller number of total friends. Such an approach will create a directed weighted graph with asymmetric weights.

We didn't provide our participants any information about how we see this method used. Still, some of the responses we received were quite interesting. One confirmed our notion about how such a method can increase privacy awareness:

> *"... scary, it brought up photos of friends that are accidently in my Facebook Is that the goal? To show that half of them are not really connected to you?"*

Another, on the other hand, surprised us with a completely new direction:

> *"... it was really fun for me, this little game ... "*

This quote highlights the potential of an interactive process for privacy policy configuration. Such a process can be formed as as series of questions, much like our experiment, that will adjust the social graph, by changing the weights, until the appropriate clustering is achieved. Following such an assisted clustering process, the user can configure the policy over the different groups.

Practical Implications

Recently (June 2010) Facebook announced it passed the 400 million active users mark, with the average user having 130 friends [8]. With the flood of information large networks create, the creation of complete and detailed network representations is becoming harder and more complex. Concise representations have obvious benefits: they are quick to digest by the users and require relatively little screen real estate. Such representations naturally contain less information, however our evaluation shows that compact representations encapsulate much more knowledge than they display.

Our vision is a social network that will keep the user constantly aware of her privacy. The importance of keeping users informed about their exposure is highlighted as more and more personal information migrates to social networks. An obvious way to handle this challenge is to list all the users that have access to a certain item along it. Such detailed representation of privacy is neither possible nor effective. A compact representation of the social network, as we are suggesting, can answer this challenge by offering a settlement between the accurate and detailed representation and usefulness.

In general we show that minimal representations, when leveraging the user's knowledge, can be quite effective. This is something that social network designers and developers in general should take into account.

FUTURE WORK

Adding More Tie Features

In Table 1 we list features that could improve the predictive capabilities of our clustering platform. In future work we believe it important to evaluate the power of incorporating

these features into a clustering system. Most of these features are unavailable through the Facebook API, and therefore not measurable using our current system. Therefore, future research might involve novel ways of performing such research that does not rely on the specific API of an existing system. Such features include communication features, such as the number of "wall" words exchanged, network features, such as the size and sparseness of each user's network, and personal information, such as age.

Fitting the Representation Generation to User's Network

In addition to personal data, the properties of a user's network - such as path lengths, clustering coefficient, etc' - might also have an affect on how their social graph is best clustered. Such properties can influence the choice of the community detection algorithm, number of clusters and representing members detection methods. Future work can utilize such information to better select the methods used and evaluate the representation.

Experimenting ''In the wild''.

Our study evaluates instance-based representation as a general technique for visualizing a social graph. It does not, however, study the particular applications that provided our motivation to this technique. In future work, we hope to deploy a privacy related application that uses instance-based representation to communicate which members of a user's social graph can access individual resources. It's also possible to compare this representation to existing solutions in Facebook, as seen in figure 1.

Privacy Policy Configuration

Creating useful user interfaces for manipulating complex security settings is an open problem. We think it's possible to leverage a similar platform to the one we used to conduct our study to assist users in configuring privacy settings. Specifically, a good clustering algorithm could help alleviate the problem of dividing a social graph into related groups by suggesting closely related friends to the user. In fact, while conducting our study, several of the participants found it fun to participate, and viewed the platform as a sort of *game* over their social graph. We think a similar platform geared at *setting* privacy policies would make this task less daunting to the user.

REFERENCES

1. Privacy groups challenge Facebook on new settings. `http://news.bbc.co.uk/2/hi/technology/8420431.stm`, 2009.
2. E. Baatarjav, R. Dantu, and S. Phithakkitnukoon. Privacy management for Facebook. *Information Systems Security*, pages 273–286.
3. Facebook inc. `http://www.facebook.com`.
4. G. Gates. Facebook Privacy: A Bewildering Tangle of Options. `http://www.nytimes.com/interactive/2010/05/12/business/facebook-privacy.html`, 2010.
5. E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 211–220. ACM New York, NY, USA, 2009.
6. M. Granovetter. The strength of weak ties. *ajs*, 78(6):1360, 1973.
7. F. inc. Facebook developers. `http://developers.facebook.com/`, 2010.
8. F. inc. Facebook statistics. `http://www.facebook.com/press/info.php?statistics`, 2010.
9. E. Lieberman and R. Miller. Facemail: showing faces of recipients to prevent misdirected email. In *Proceedings of the 3rd symposium on Usable privacy and security*, page 131. ACM, 2007.
10. H. Lipford, A. Besmer, and J. Watson. Understanding privacy settings in facebook with an audience view. *Usability, Psychology, and Security*, 2008.
11. D. McCullag. House panel presses Facebook, Google on privacy. `http://edition.cnn.com/2010/TECH/social.media/06/01/cnet.congress.facebook.privacy/index.html`, 2010.
12. L. Palen and P. Dourish. Unpacking privacy for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, page 136. ACM, 2003.
13. P. Pons and M. Latapy. Computing communities in large networks using random walks. *Computer and Information Sciences-ISCIS 2005*, pages 284–293, 2005.
14. M. Porter, J. Onnela, and P. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 2009.
15. B. Rohan. German minister sees Facebook fined over privacy. `http://www.cnbc.com/id/37493255`, 2010.
16. Walktrap v0.2. `http://www-rp.lip6.fr/~latapy/PP/walktrap.html`.