You Are Who You Know: Inferring User Profiles in Online Social Networks

Alan Mislove[†‡§], Bimal Viswanath[†], Krishna P. Gummadi[†], Peter Druschel[†]
[†]MPI-SWS, Saarbrücken, Germany
[‡]Rice University, Houston, TX, USA
[§]Northeastern University, Boston, MA, USA
amislove@ccs.neu.edu, {bviswana, gummadi, druschel}@mpi-sws.org

ABSTRACT

Online social networks are now a popular way for users to connect, express themselves, and share content. Users in today's online social networks often post a profile, consisting of attributes like geographic location, interests, and schools attended. Such profile information is used on the sites as a basis for grouping users, for sharing content, and for suggesting users who may benefit from interaction. However, in practice, not all users provide these attributes.

In this paper, we ask the question: given attributes for some fraction of the users in an online social network, can we *infer* the attributes of the remaining users? In other words, can the attributes of users, in combination with the social network graph, be used to predict the attributes of another user in the network? To answer this question, we gather fine-grained data from two social networks and try to infer user profile attributes. We find that users with common attributes are more likely to be friends and often form dense communities, and we propose a method of inferring user attributes that is inspired by previous approaches to detecting communities in social networks. Our results show that certain user attributes can be inferred with high accuracy when given information on as little as 20% of the users.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Human factors, Measurement

Keywords

Social networks, inferring attributes, communities

1. INTRODUCTION

Online social networks are now a popular way for users to connect, express themselves, and share content. For example, MySpace (over 275 million users)[1], Facebook (over 300 million users), Orkut (over 67 million users), and LinkedIn (over 50 million "professionals") are examples of wildly popular networks used to find and organize contacts. Some networks such as Flickr, YouTube, and Picasa are used to share multimedia content, and others like LiveJournal and BlogSpot are popular networks for sharing blogs.

Users often post profiles to today's online social networks, consisting of *attributes* like geographic location, interests, and schools attended. Such profile information is used as a basis for grouping users, for sharing content, and for recommending or introducing people who would likely benefit from direct interaction. Today's online social networks rely on users to manually input profile attributes, representing a significant burden on users, especially when users are members of multiple online social networks. Thus, in practice, not all users provide these attributes, thereby reducing the usefulness of the social networking applications.

In this paper, we ask the question: is it possible to *infer* the missing attributes of a user in an online social network from the attribute information provided by other users in the network? In other words, can the attributes of other users in the network, in combination with the social network graph, be used to predict those of a given user? In offline social networks, people often socialize with others who share the same interests, geographic location, or alma mater. Thus, it is natural to try to leverage the attributes provided by users in order to predict those of their friends. The ability to automatically predict user attributes could be useful for a variety of social networking applications such as friend and content recommendations, and scoped content sharing. On the other hand, answering this question has important privacy implications, as a user's privacy may no longer depend only on what he or she reveals to the various social networks.

To answer this question, we collect two detailed social network data sets. Our first data set covers the social network of almost 4,000 students and alumni of Rice University collected from Facebook [7]. For each student, we gather attributes like major(s) of study, year of matriculation, and dormitory, to see if these attributes can be inferred from friends in the social network. Our second data set covers over 63,000 users in the New Orleans Facebook regional network. For each user in this data set, we also collected their profile page, which lists a large number of user-provided attributes. For both data sets, we find that users are significantly more likely to be friends with users with similar

[1]The number of users refers to the number of identities as published by the social networking sites in November 2009.

WSDM'10, February 4–6, 2010, New York City, New York, USA.

attributes, and that groups of users with common attributes often form dense subgraphs.

We propose a new approach for inferring the attributes of users. Inspired by existing work on community detection, we start with a seed set of users with known attributes and look for communities of users in the network based around this seed set. As a community is generally defined as a group of users who are more tightly interconnected than the surrounding graph, detecting communities that are centered around users with a common attribute is a natural approach to predicting other users that share the attribute. Our results show that this approach works surprisingly well: depending on the strength of the community in the network, user attributes can often be inferred with high accuracy when given information about as few as 20% of the users. For example, in our data set we can, with high accuracy, predict user attributes such as matriculation year, dormitory, and high school.

The rest of this paper is organized as follows. Section 2 describes the social network data we collected and its limitations. Section 3 examines our collected data and demonstrates that the structure of the social network correlates well with user attributes. Section 4 details our approach for inferring attributes, and presents an evaluation on real-world social network data. Section 5 discusses related work and Section 6 concludes.

2. DATA COLLECTED

In this section, we describe each of the data sets we collected and their limitations.

2.1 Rice University data set

Our first data set is the Rice University Facebook network.

2.1.1 Measurement methodology

This data set was collected by crawling part of Facebook [7] through the site's public web interface. We crawled the Rice University Facebook network, which consists of Rice University students and alumni. We started by logging into the Facebook user account of one of the authors, who is a student at Rice University. We then conducted a breadth-first-search (BFS) of all reachable users in the Rice network, in the same manner as in previous work [15]. By default, Facebook allows all users in the same network to view each others' friends, and we were thus able to crawl a large portion of the Rice Facebook network.

The data collected for this paper is from a crawl conducted over 9 hours on May 17th, 2008. In total, our crawl discovered 6,156 users, who are connected together with 188,675 undirected links. This represents an average degree of 61.29.

2.1.2 User attributes

From the Facebook crawl, we were only able to collect the name of the users and their list of friends. We collected additional information about the users by querying the Rice University Student Directory [22] and the Rice University Alumni Directory [21]. From these two directories, we were able to determine the users' matriculation year, graduation year, residential college[2], and major(s) or department.

To correlate the Facebook user list with the directories, we first looked up each user's name in the Student Directory, and then the Alumni Directory. If a single entry was found in either directory, the information from that entry was used.[3] If multiple entries were found that exactly matched the student's name, we disregarded the student. We used a conservative matching policy: only exact name matches (with allowances for common nicknames) were used.

Overall, we found matches for 1,781 students in the Student Directory and 2,093 additional students in the Alumni Directory.[4] This left us with 2,282 Facebook users who we were unable to match with a directory listing; we disregarded these users. Of the 3,874 students we were able to find records for, 1,220 (31.5%) were current undergraduate students, 501 (12.9%) were current graduate students, 1,856 (47.9%) were undergraduate alumni, and 237 (6.11%) were graduate alumni. The total number of current undergraduate and graduate students at Rice is 3,001 and 2,144, respectively [20]. Thus, we were able to locate 40.7% of the current undergraduate and 23.4% of the current graduate students in Facebook.

2.1.3 Data sets used

Throughout the next few sections, we consider two subsets of the Rice data set representing different parts of the Rice University network. The first subset we use is the current undergraduates. This subset contains 1,220 users connected with 43,208 undirected links, for an average degree of 70.8.[5] The second subset we use is the current graduate students. This subset contains 501 users connected with 3,255 undirected links, for an average degree of 12.9. We examine these two parts of the network separately, since we have different attributes sets for the undergraduates and graduate students and they represent largely distinct parts of the network. In fact, only 1,395 links (2.9% of all links) are present between the undergraduate and graduate networks.

2.2 New Orleans data set

Our second data set is the New Orleans Facebook network.

2.2.1 Measurement methodology

We collected this data set largely in the same manner as the Rice data set, starting with a seed user and crawling using a breadth-first search. Facebook allows any user to join regional networks, so we were able to create multiple accounts for crawling in the New Orleans network in parallel. The data was collected over a five day period starting on December 29th, 2008. In total, using the same crawling methodology as above, we discovered 90,269 users connected by 1,823,331 undirected links, for an average degree of 40.39.

[2]Rice University has nine residential colleges, to which incoming undergraduate students are randomly assigned. The colleges serve as dormitories, cafeterias, and social circles; students stay at the same college during their entire undergraduate tenure.

[3]The only exception was for alumni who graduated before 1980; such users are unlikely to have Rice University email accounts, and are therefore unlikely to have accounts in the Rice University Facebook network. As a result, we disregarded these matches.

[4]Note that Rice students can elect to remove their information from the online directory; in this case, we would not be able to find corresponding entries in the directories.

[5]Our average user degree is lower than is cited by Facebook at `http://www.facebook.com/press/info.php?statistics` since we only have intra-Rice links. Links to other accounts not in the Rice network are not included.

2.2.2 User attributes

In order to collect attributes for users in the New Orleans network, we also collected the user profiles during the crawl. Each profile consists of optional information provided by the users themselves, such as educational information, tastes and preferences, and geographic information. Since users are allowed to mark their profiles as private, we were not able to download profile information for all users. In total, we were able to download profiles for 63,731 (70.6%) of the users, and we consider only this subset in the following analysis.

Attribute	Fraction revealed
high school	68.9%
university	58.3%
employer	42.3%
interests	35.5%
location	19.3%

Table 1: Fraction of users who provide various attributes in the New Orleans Facebook network.

We also conducted a quick study to determine what fraction of users provide various attributes in their Facebook profiles. Table 1 lists the fraction of users who provide different attributes in their profile in the New Orleans network. The rates for different attributes vary widely: for example, almost 70% of users provide their high school, but only 20% of users provide their current city of residence. This observation shows that automatically inferring user attributes could be useful to today's online social networks.

2.3 Limitations

Both of our Facebook crawls include only those users who had not changed the default Facebook privacy settings, which shares their profile and friend list with users in the same network. During our crawls, we found that about 5% of each network had changed their privacy settings so that their friend list was inaccessible, and about 30% of the network had made their profile inaccessible.

Additionally, we may have missed users who were not connected to the large, strongly connected component of the social networks we crawled. Because Facebook does not provide a way to select random users, we are unable to estimate the fraction of accounts that we were unable to crawl.

3. ATTRIBUTES IN THE NETWORK

Our approach to inferring user attributes is based on two observations about how the structure of the social network is correlated with the attributes of users. First, we note that users are significantly more likely to be friends with other users who share their attributes. In some cases, the likelihood is as high as 53-fold more than what would be expected if attributes were assigned randomly. Second, we observe that this tendency for similar users to be linked often leads to *communities* of users in the network that are centered around attributes. Each of these observations are described in more detail below.

3.1 Friends with common attributes

Our first observation is that users are statistically much more likely to be friends with other users who share their attributes, when compared to users who have no attributes in common. In order to show this, for each attribute a (such as college, matriculation year, or high school), we calculated

$$S_a = \frac{|\{(i,j) \in E : \text{ s.t. } a_i = a_j\}|}{|E|} \quad (1)$$

where a_i represents the value of attribute a for user i, and E represents the set of all links. S_a therefore represents the fraction of links for which users share the same value of attribute a. We divided this by E_a, or what would be expected if attributes were placed randomly,

$$E_a = \frac{\sum_{i=0}^{k} T_i(T_i - 1)}{|U|(|U| - 1)} \quad (2)$$

where T_i are the number of users with each of the possible k attribute values and $U = \sum_{i=0}^{k} T_i$. The resulting value $A_a = S_a/E_a$, which we call *affinity*[6], ranges from 0 to ∞ and represents the ratio of the fraction of links between attribute-sharing users, relative to what would be expected if attributes were assigned randomly. Thus, an affinity greater than 1 indicates that links are positively correlated with user attributes.

Users	Attribute	Affinity
Rice undergrads	college	4.49
	major	2.33
	year	1.97
Rice grads	department	9.71
	school	4.02
	year	1.79
New Orleans	high school	53.2
	hometown	2.87
	political views	1.86

Table 2: Affinity values for various attributes. Links are correlated with numerous user attributes.

Table 2 shows the affinity of the various attributes for our crawled data sets. We observe that for a number of the attributes, a significant affinity is observed, showing that links are correlated with certain attributes. It is interesting to note that certain attributes have stronger affinity than others: for example, graduate students have a much strong affinity for other students in the same department than to other students in the same matriculation year. For some attributes, the affinity is as high as 53, implying that users connected by a link are 53 times more likely to share an attribute than would be expected if attributes were random. In summary, we have observed that links are correlated with certain attributes, suggesting that our approach of inferring attributes from the social network structure holds promise.

3.2 Attribute-based communities

Given that we have observed a correlation between user attributes and links, it is natural to see if the users who share a similar attribute form communities, or dense clusters, in the network. Note that the previous observation is a necessary, but not sufficient, condition for attribute-based communities to exist. For example, users linked by a common attribute could form a long chain, having high affinity but not forming a dense community. In order to investigate whether attribute communities are present in our network, we divide the network into communities based on user

[6]Affinity essentially represents the degree of homophily in the network, with respect to a particular attribute.

attributes, and then quantify the strength of the resulting communities using modularity [17].

3.2.1 Modularity

Consider a partitioning of a network into k distinct communities. Let $\mathbf{e}$ be a symmetric $k \times k$ matrix, whose element e_{ij} is the fraction of edges in the network that connect vertices in community i to community j. Also, we define $a_i = \sum_j e_{ij}$ as the fraction of edges that touch vertices in community i. Then, the trace of the matrix Tr $\mathbf{e} = \sum_i e_{ii}$ gives the fraction of edges in the network within the same community. Hence, modularity is defined as

$$Q = \sum_i (e_{ii} - a_i^2) = \text{Tr } \mathbf{e} - ||\mathbf{e}^2|| \quad (3)$$

where $||\mathbf{y}||$ indicates the sum of the elements of matrix $\mathbf{y}$. Modularity is then a measure of the fraction of intra-community edges minus the expected value of the same quantity in a network with the same community divisions, but with edges placed without regard to communities. Modularity therefore ranges from -1 to 1, with 0 representing no more community structure than would be expected in a random graph, and significantly positive values representing the presence of strong community structure.

3.2.2 Rice undergraduates

Table 3 shows the modularity for the undergraduate population when partitioned according to residential college, major, and matriculation year. Also shown is the modularity of the partitionings that are obtained when multiple attributes are used. The results show a significant modularity for the communities defined by residential college and matriculation year - a relatively high Q of 0.384 is observed when partitioning by residential college, and a Q of 0.259 is seen when dividing by year. However, the modularity of the communities defined by major is almost 0, indicating that no community structure exists based on academic major. Overall, these results indicate that undergraduates who share the same college or matriculation year form tightly-knit communities in the social network.

Attributes	Communities	Modularity
college, major, year	582	0.023
college, major	317	0.029
year, major	147	0.045
major	52	0.055
college, year	44	0.248
year	7	0.259
college	9	0.384

Table 3: Modularity values for attribute communities for undergraduates at Rice. College and matriculation year reveal strong community structure.

With some knowledge of the actual social network at Rice, the above results are not unexpected. Undergraduate students are randomly assigned to a residential college upon matriculation, and they generally remain members of that college for the duration of their undergraduate studies. Thus, it is natural that strong communities form around residential colleges. Additionally, the strong communities among undergraduate students of the same matriculation year are not surprising. Incoming students attend an orientation week together, are mostly assigned to share dormitory rooms with students of their year, and tend to spend time in courses with students of their year. Thus, it is also natural that a community structure exists among undergraduates of the same matriculation year. Finally, the lack of a strong community structure around majors can be explained by the fact that Rice undergraduates obtain a liberal arts education (taking courses from many departments), and they often do not choose majors until the end of their sophomore year.

3.2.3 Rice graduate students

We now turn our focus to the graduate student population. Table 4 shows the modularity of the graduate student population when partitioned according to department, academic school, and matriculation year.[7] The results show a significant modularity for the communities based on department – in fact, a Q of 0.587 is observed. A similar modularity value is observed when partitioning according to school – this is because each department is a member of exactly one school, and the partitioning according to school ends up being a coarser version of the communities defined by department. Finally, a Q of 0.185 is seen for the communities defined by matriculation year. This indicates a very strong community structure for the graduate students based on department, and a weak community structure based on matriculation year.

Attributes	Communities	Modularity
year	10	0.185
department, school, year	124	0.292
department, year	124	0.292
school, year	43	0.299
school	7	0.581
department, school	28	0.587
department	28	0.587

Table 4: Modularity values for attribute communities for graduate students at Rice. Departments form strong communities.

The results for the graduate student population are also not unexpected. Graduate students are accepted into a specific department at the beginning of their studies, and usually spend their entire tenure in the same department. Thus, the very strong association with the department is not surprising. Moreover, the variable length of graduate programs and the greater tendency of graduate students to interact across seniority levels explains why the partitioning according to matriculation year has a weak community structure.

For brevity, we do not include results in this section for the New Orleans network, however, we obtained similar results for attributes like high school and hometown.

3.3 Summary

In all three of our data sets, we observe that users with certain similar attributes tend to be friends in the social network. Moreover, we observe strong communities, indicated by a high modularity value, for the communities defined by users who share certain attributes in the Rice networks. We also observe that multiple overlapping community structures exist. For the undergraduates, we observe significant modularity when partitioning according to residential college and matriculation year. For the graduate students, we observe significant modularity when partitioning according

[7]Note that graduate students are not assigned to residential colleges, so that attribute is disregarded here.

to department and weaker modularity when partitioning by matriculation year.

4. INFERRING ATTRIBUTES

In the previous section, we used knowledge of all attributes in the network to examine the communities defined by users who share attributes. In this section, we examine the problem of detecting these communities even if we don't know all of the attributes. Our approach is based on the observation that strong community structures often exist around users with common attributes. This observation suggests a natural way of inferring user attributes if the attributes for some users are not known: namely, to infer user attributes by detecting communities in the network. In this section, we describe our approach and results. We first describe related work on community detection that we leverage to infer attributes, and then present an evaluation on our Rice and New Orleans data sets.

4.1 Community detection

Community detection in large networks is a well-studied problem with a number of notable approaches. At a high level, algorithms for detecting communities can be divided into *global* approaches, which assume knowledge of the entire network, and *local* approaches, which only assume knowledge of a local region. We briefly discuss each of these below.

4.1.1 Global community detection

One of the first community detection algorithms was proposed by Girvan and Newman [18]. Their algorithm works by iteratively removing edges until the social network graph becomes partitioned, at which point the various partitions are considered communities. In order to determine the edge to be removed at each step, Girvan and Newman proposed a metric known as *betweenness centrality* for each edge. To compute this metric, it is necessary to compute the shortest path between each pair of vertices in the network. The number of shortest paths that contain an edge determine the betweenness centrality of that edge. Follow-up work has extended the approach taken by Girvan and Newman in various ways, with significant speed improvements [17, 19, 23].

The intuition behind this algorithm is simple. If we assume that the social network is divided into densely connected communities, the betweenness centrality metric looks for links that bridge communities. Since communities are, by definition, more dense than the graph as a whole, these bridging links will naturally have a higher betweenness centrality. Once they are removed from the graph, the underlying community structure emerges.

4.1.2 Local community detection

One potential downside of the global approaches to community detection is that the structure of the entire graph must be known; as others have pointed out [4], this is often prohibitively expensive (as many real-world graphs are extremely large) or hard to obtain (for example, the graph of Web pages). As an alternative, a number of researchers have looked at local approaches to detecting communities, which use only local knowledge to build a community around a set of source nodes. In contrast with the global approaches, local approaches have the potential to be significantly more scalable and applicable to much larger graphs.

Most of the local approaches work by starting with a single (or multiple [2]) seed node and greedily adding neighboring nodes until a sufficiently strong community is found. For example, Clauset's algorithm [4] at each step adds the node that maximizes the ratio of intra-community edges to inter-community edges for the nodes on the "fringe" of the community. Bagrow's algorithm [3] adds the node which has the lowest "outwardness", which is defined as the number of neighbors outside the community minus the number within, normalized by degree. Finally, Luo et al. [13] proposed an algorithm similar to Clauset's but with the metric based on all the nodes in the community and not just the fringe. It also performs iterative add and remove cycles, iterating until adding or removing a single vertex can no longer result in a better community.

4.2 Inferring attributes globally

The first scenario we examine is whether we can infer attributes at a global scale. For example, if we know the matriculation year for 10% of the users, how well can we infer the matriculation year of the remaining 90%?

Our approach is to detect communities at a global level, seeded with the partial information about user attributes. In particular, we modified Clauset's algorithm [5] to make use of attributes of a subset of the users. Instead of starting with every user in their own cluster, the algorithm pre-assigns users with the same attribute value into the same cluster. We then run the algorithm as normal, effectively "seeding" it with the users who reveal their attributes. Finally, we compare the resulting communities with the communities based on the known attributes of all users.

To measure how similar these two community structures are, we use the *normalized mutual information* metric [9]. This metric is calculated as

$$\frac{-2\sum_i\sum_j \mathbf{x}_{ij} log(\frac{\mathbf{x}_{ij}N}{\mathbf{x}_{i.}\mathbf{x}_{.j}})}{\sum_i \mathbf{x}_{i.} log(\frac{\mathbf{x}_{i.}}{N}) + \sum_j X_{.j} log(\frac{\mathbf{x}_{.j}}{N})} \quad (4)$$

where $\mathbf{x}$ is a square matrix whose dimension is the number of communities detected. Each element $\mathbf{x}_{ij}$ represents the number of nodes in attribute-defined community i that appeared in the detected community j. The quantities $\mathbf{x}_{.i}$ and $\mathbf{x}_{i.}$ denote the sum over column i and row i respectively, and N is the number of nodes in the graph. The metric ranges between 0 and 1, with 0 representing no correlation between the two community structures, and 1 representing a perfect match.

Figure 1 plots the results of this experiment for the Rice undergraduates, by showing the normalized mutual information for each attribute. Separate lines are plotted for each attribute, and the correlation value is with respect to the attribute that users are revealing. Two trends can be seen in this graph. First, we observe that both college and year quickly lead to community structures with significant correlation. In fact, when just 20% of users reveal their college or year, we can infer the attributes for the remaining users with over 80% accuracy. Second, this is not the case for major of study. However, this result is not surprising, as we observed in the previous section that communities are not formed around users with common majors. Overall, this experiment shows that multiple attributes can be inferred globally when as few as 20% of the users reveal their attribute information.

Similarly, Figure 2 plots the results of this experiment for

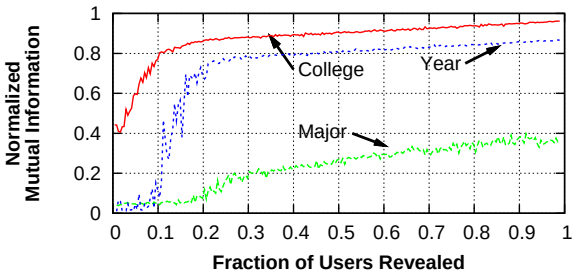


Figure 1: Normalized mutual information versus the fraction of users who reveal their community for Rice undergraduates. Revealing more information naturally leads to partitionings with higher correlations, especially for the college and year attributes. This result shows that different attributes can be accurately inferred with as few as 20% of users revealing their attributes.

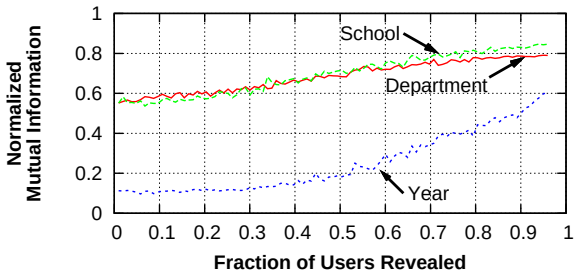


Figure 2: Normalized mutual information versus the fraction of users who reveal their community for Rice graduate students.

the Rice graduate students. Similar to the undergrads, we observe that certain attributes correspond to communities that can be detected with high accuracy. For example, if as few as 5% of the students reveal their department or school, we can infer the department or school for the remaining students with approximately 60% accuracy. However, this is not the case for the matriculation year attribute. We observed in the previous section that matriculation years form only weak communities, so this result is not unexpected.

We were unable to conduct this experiment for the New Orleans network, since the attributes in that network are self-reported, and there are not any attribute types that are known for all users.

4.3 Inferring attributes locally

We now look at detecting attributes on a local scale. This is different from the problem in the previous section, where we assumed that partial information is known about all attribute values. Instead, for example, we may know that a subset of five users all live in the same dormitory, and we wish to determine the other users (for which we do not have any information) who also live in that dormitory. To detect these communities, we extend the previously proposed approaches for local community detection to take a seed set.

While exploring local community detection, we found that previous approaches performed well when detecting certain attributes, but did not perform well on others. For example, as we examine later, we found that the algorithm of Luo et al. [13] could infer the undergraduate members of a residential college at Rice, but was not able to infer the members of weaker communities, such as all students in the same matriculation year. Thus, we propose a new method for detecting a single community, based on the metric of *normalized conductance*. We first describe this new metric below, then give a description of our algorithm, and finally evaluate the algorithm on our data sets.

4.3.1 Normalized conductance

We first define a metric that rates the quality of a single community (as opposed to modularity, which rates the community structure of a partitioning of a graph into a collection of communities). To measure the quality of a community, we propose a metric based on the widely adopted metric conductance [12]. Let $G = (V, E)$ denote a graph, let $A \subset V$ be a subset of the vertices that forms a community, and let $B = V \setminus A$. Let us also define e_{AB} to be the number of edges between A and B and e_{AA} as the number of edges within A. The conductance of A is then traditionally defined as e_{AB}/e_{AA}. Therefore, a small value of conductance denotes a strong community, as the community would be tightly linked internally, with very few external links.

However, this definition of conductance is not a good measure for the quality of a community, as it is biased towards large communities. For example, if we place all vertices in a single community, the conductance would be 0, providing little information about the community formed.

Hence, we propose a new metric called *normalized conductance*. To derive normalized conductance, we first define the value K of community A as

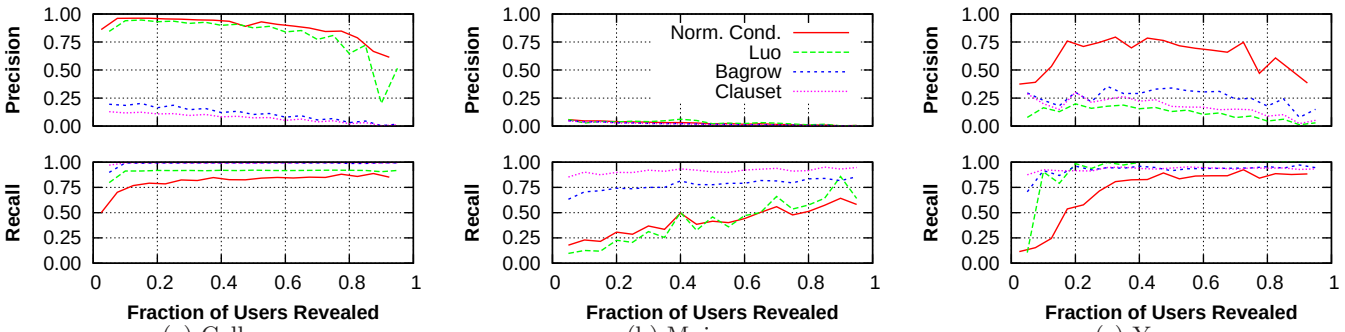
$$K = \frac{e_{AA}}{e_{AA} + e_{AB}} \tag{5}$$

This value is similar to conductance, except that it ranges between 0 and 1. A measure close to zero indicates very poor community structure, and a measure close to 1 indicates very good community structure with many more links within A than to the outside. However, this metric is still not perfect, as very large communities are naturally biased towards having many more edges within the graph (high e_{AA}). Thus, we define the normalized conductance C for a community A as K minus the expected value of K for a random graph divided into communities of sizes $|A|$ and $|B|$.

To calculate the expected value of K for a random graph, we need to calculate the expected values of e_{AA} and e_{AB} for a graph with the same community division and degree distribution, but with the links placed without regard to the communities. We define $e_A = e_{AA} + e_{AB}$ and $e_B = e_{BB} + e_{AB}$, with e_A denoting the number of edges that reach vertices within A, and e_B giving the same quantity for B. In a random graph, we would expect that $e_{XY} = e_X e_Y$. Thus, our normalized conductance metric C can be written as

$$C = \frac{e_{AA}}{e_{AA} + e_{AB}} - \frac{e_A e_A}{e_A e_A + e_A e_B} \tag{6}$$

The metric C ranges between -1 and 1. Similar to modularity, strongly positive values indicate significant community structure in A, a value of 0 indicates no more community structure than a random graph, and strongly negative values indicate less community structure than a random graph. One particularly useful property of this definition of conductance is that it is comparable across graphs of different sizes and densities.



(a) College (b) Major (c) Year

Figure 3: Average recall and precision of single community detection for Rice undergraduates for multiple algorithms. Good performance is observed for our algorithm (Norm. Cond.) for college and year; detection of users with the same major is poor due to the low correlation with communities in the network. The algorithm of Luo et al. performs well at inferring college but does not perform well for inferring matriculation year.

4.3.2 Algorithm

We now describe our algorithm for detecting a single community, using the normalized conductance metric C. We assume the algorithm is given as input a subset of users S in a community and the social network graph $G = (V, E)$. The algorithm then returns the other members of the community. Similar to the approach that was taken by Luo et al. [13], we use a greedy approach to maximize the normalized conductance. We divide the graph into two components A and B, with $A = S$ initially. At each step, we select a user $v \in V$ in B that upon adding v to A yields the highest increase in the normalized conductance C for A. We repeat this process, adding users to A, until no remaining user would produce an increase in the normalized conductance C for A. At this point, we stop and return the community A as the result.

The primary difference between our method and the previous approaches is the use of a metric that is weighted against a random graph. We found that the metrics used by previous approaches are all biased towards large communities. For example, the metric used by Luo et al. [13] is based on the ratio between the number of intra-community links to the number of inter-community links. As a community grows larger, this value naturally increases; in fact, it becomes infinite if an entire connected component is viewed as a community. Thus, these approaches often have trouble detecting large communities in the network, as once the community is detected the algorithm does not stop, but continues to add nodes to the community until the community is defined as the entire graph. By weighting our metric against a random graph, we can detect both the small-scale and large-scale communities that exist.

4.3.3 Evaluation

To see how well our algorithm and others perform, we evaluate the performance along two axes. Assume that each algorithm takes as input a subset S of the set H of users with a certain shared attribute, and the social network graph. The algorithm then returns a set of users R, representing the other members it believes also belong in H, based on the community structure in the network. We define the *recall* to be $|R \cap H|/|H \setminus S|$ representing the fraction of remaining community members returned. Similarly, we define the *precision* to be $|R \cap H|/|R|$ representing the fraction of the returned users who are actually in the community. Thus, an ideal algorithm would have a recall of 1 (returning all of the remaining users) as well as a precision of 1 (only returning users who are actually in the community).

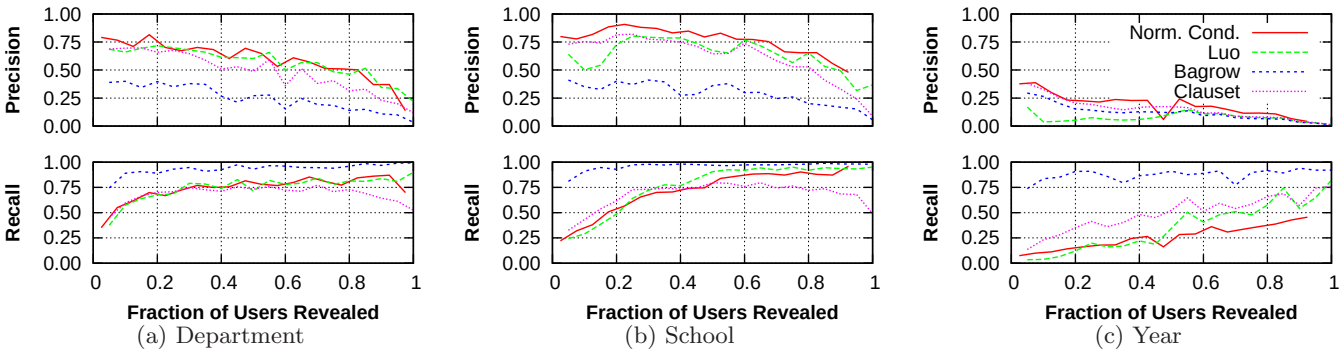
We now evaluate our algorithm along with the algorithms of Luo et al. [13], Bagrow [3], and Clauset [4]. First, we examine how well they perform on the undergraduate population by providing the algorithms with varying-size subsets of the students with common attributes such as college, matriculation year, and major. For each attribute (i.e., each college, each major), we select 20 random subsets of users of each size. We then evaluate how well the algorithms perform when given each of these random subsets as input.

For fair comparison with the other algorithms, a few parameters and modifications were required. First, none of the other algorithms accept as input a set of seed nodes; we naturally extended them to start with a set of nodes rather than a single node. Second, the algorithm proposed by Clauset does not specify a stopping condition; instead, it requires the user to specify the number of nodes to be added to the community. Thus, we utilize the stopping condition proposed by Bagrow [3] for the Clauset algorithm, based on *p-strong* communities.[8] We ran the algorithms of Clauset and Bagrow with values of $p = \{0.75, 0.8, 0.85, .., 1.0\}$, as suggested, and selected the one with the lowest number of inter-community edges (representing the "best" community). Third, the algorithm of Lou et al. performs iterative additions and deletions, and could therefore remove the original seed nodes from the resulting community. In order to handle this case, we imposed the constraint that we only consider the algorithm of Luo to have found a community if 50% or more of the original seed nodes were present in the resulting community.

4.3.4 Inferring attributes for Rice undergrads

We now present the results for inferring different attributes for the Rice undergraduate students. For these results, we average over all possible values of each attribute (such as all colleges) in order to compute the recall and precision data presented in Figure 3. Thus, we feed each algorithm x% of every college and calculate the recall and precision of the result. We repeat this experiment 20 times

[8] A community is p-strong when a fraction p of nodes within the community satisfy the criteria that they have more neighbors inside the community than outside



(a) Department (b) School (c) Year

Figure 4: Average recall and precision for single community detection for Rice graduate students. Good performance is observed for department and school; much weaker performance is seen for year.

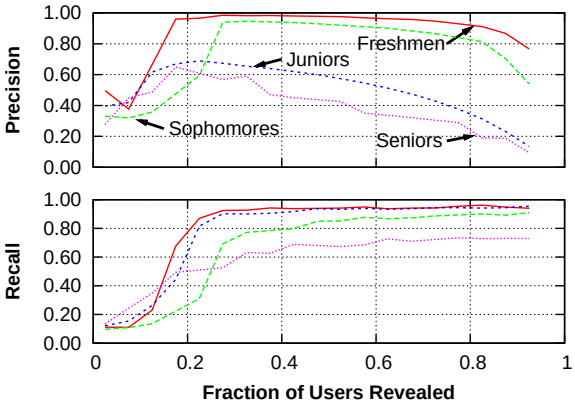


Figure 5: Recall and precision for matriculation year community detection for Rice undergraduates for our algorithm. Individual lines are shown for each matriculation year. Performance depends on the matriculation year.

for each college and fraction revealed, and then average over all colleges to obtain the data in Figure 3(a).

As a detailed example, Figure 5 presents the recall and precision for each of the matriculation years as different number of users are revealed. A number of interesting observations can be made about the results. First, the performance varies across the different matriculation years; the freshmen and sophomores appear to be the easiest to detect, followed by the juniors and seniors. Second, detection performance is good for all of the matriculation years once 20% to 30% of the users are revealed. Third, note that the precision naturally deteriorates once very high fractions of the users in each year are revealed. This is because the precision is defined based on the number of unrevealed users, which becomes much smaller as significant fractions are revealed. We now turn back to Figure 3 and discuss each attribute in detail.

Colleges: The results show that colleges can be inferred with very high recall and precision by both our algorithm and the algorithm of Luo et al. when as few as 20% of the students in the college are known. For example, when 20% of the members of a single college are provided to the algorithms, both our algorithm at that of Luo et al. can infer over 80% of the remaining members of that college with over 95% accuracy. The algorithms of Clauset and Bagrow both perform rather poorly at detecting colleges: they each often identify a large part of the network as belonging to the college, resulting in a very low precision score.

Years: However, for inferring matriculation years, all algorithms have high recall, but only our algorithm has good precision. In fact, the other algorithms tend to detect the entire graph as a community, which leads to the low precision. Again, we believe that this poor performance is a function of the metrics that the other algorithms use. Since they essentially try to maximize the ratio of intra-community links to inter-community links, they occasionally end up returning the whole graph.

Majors: Finally, we observe that none of the algorithms are able to infer major; all have extremely low precision. This result is expected, though, since we observed in the previous section that majors do not form significant communities in the network.

4.3.5 *Inferring attributes for Rice graduate students*

We now evaluate our approach on the Rice graduate student network. Figure 4 shows how the recall and precision vary as different fractions of the department, school, and matriculation year of graduate students are provided. All algorithms perform well when inferring the department and school, with the exception of Bagrow's. (As we observed with the undergraduates, the algorithm of Bagrow tended to return a large portion of the network as a community.) We find that knowing 20% of the user attributes is sufficient to infer most of the remaining users with high accuracy. However, none of the algorithms perform well at inferring matriculation year. Again, the poor performance at detecting matriculation years can be explained by the data in Section 3, which shows that the students with the same matriculation year form weak communities in the social network.

4.3.6 *Inferring attributes for New Orleans users*

In the last section, we evaluate our technique's ability to infer University-related attributes, obtained not from Facebook from the Rice student directory. Here, we evaluate our technique's ability to infer self-reported, non-authoritative attributes. To do so, we use our New Orleans Facebook data set, which consists of the profiles and social network for 63,731 users in the New Orleans Facebook regional network. For our evaluation, we use the largest connected component

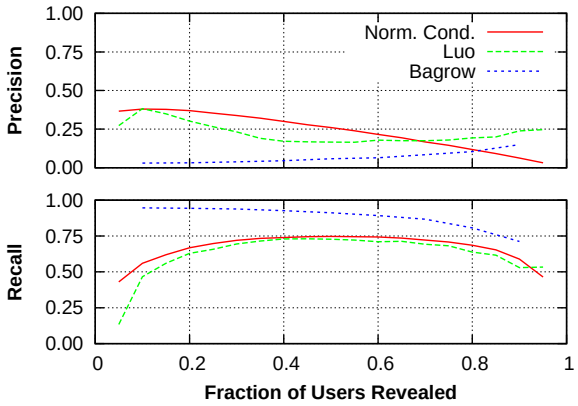


Figure 6: Average precision and recall for the 92 New Orleans groups with significant normalized conductance (> 0.2).

in this network which includes 63,392 users connected by 816,886 undirected links. Since this data was derived from user input, rather than an authoritative organization (as with the Rice data), some attributes are missing and a single attribute can be recorded in different ways by different users. Regardless, this evaluation represents a challenging environment in which to test our approach.

To evaluate how well community-based attribute inference works, we focus our evaluation on two questions: First, how many attribute-based groups represent strong communities (and could therefore possibly be inferred)? And second, for those groups, how well does our approach work? In order to answer the first question, we extracted all attributes from each user's profile.[9] In total, this resulted in 1,592,312 attributes for 63,392 users. We then examined all groups of users defined by a common attribute, looking for groups with significant conductance. In order to have sufficient granularity when selectively revealing users, we only considered groups which contained at least 15 members; in total, this represents 7,203 groups.

For each of these attribute-defined groups, we calculated their normalized conductance. Clearly, groups with low conductance are unable to be detected by our approach (as there is not information in the social network about the group), while groups with significant normalized conductance hold the potential to be detected. While most of the groups show almost no normalized conductance, there are a number of groups with significant normalized conductance. In fact, 92 (1.2%) of the groups show normalized conductance over 0.2, and 18 (0.24%) of the groups show normalized conductance over 0.3. This may seem like a surprisingly small fraction of the groups in the network, however, many of the groups with low normalized conductance are ones that would not be expected to form dense communities (examples include groups defined by common attributes like "sex: female", "birthday: december 15", and "favorite movie: Where the Wild Things Are").

Next, we attempt to detect each of these groups when given a subset of the membership. Figure 6 shows how the recall and precision vary as different fractions of each group are revealed. We were unable to run the Clauset algorithm, as the size of the New Orleans network made this algorithm computationally infeasible. While the results show worse performance than the Rice data sets, our algorithm still returns useful data. For example, with 25% of the users revealed, the normalized conductance-based approach can infer approximately 75% of the remaining users with 35% precision. The underlying reason for the lower performance is that these results essentially represent the least-favorable evaluation for the algorithm: since many of the users do not provide attributes, we are unable to tell whether the algorithm is correct or not for many of the users that are returned. To be conservative, we count such unknown users as incorrect, even though they may share an attribute but decline to list it in their profile. For example, if the group represents a high school, the algorithm may return users who do not have any high schools listed in their profile — we count these users as incorrect, even though some of these users may have attended the predicted high school.

[9]For attributes with multiple values, such as "favorite books", we treated each individual item as a separate attribute.

4.4 Summary

We began this section by asking whether we can infer user attributes if given a social network and partial information about user attributes. We demonstrated that existing techniques can be used to detect user attributes when given partial information on all attribute values and extended to accept a "seed" set of users. In fact, we found that with as few as 20% of users with known attributes, the remaining users' attributes can be inferred with over 80% accuracy. Moreover, inspired by previous work on community detection, we proposed a new algorithm to infer user attributes when provided a set of seed users who share a single attribute. In our collected networks, we found that this algorithm is able to infer multiple attributes with high accuracy when given a few users with a common attribute.

5. RELATED WORK

Before concluding, we briefly describe related work. There has been much work on automatic community detection - we provided a survey of these techniques in Section 4.1. Below, we detail other work that is related to inferring information about users from a social network.

Other studies have found that people tend to befriend others who share similar traits. In sociology, this tendency is known as homophily [16]. A study by Fiore et al. [8] of interactions of a large number of users in an online dating system showed that users usually prefer to date people who share similar attributes. Our data from Facebook agrees with these observations. In fact, homophily has been exploited to build services, for example, Şimşek and Jensen have proposed a technique [6] for navigating messages in a network by exploiting homophily.

Additionally, there have been efforts to leverage the communities formed by users who share attributes. Friendlen and Jensen [10] have proposed a family of algorithms to detect *tribes* or groups of individuals who are tightly linked to each other in an anomalous way (meaning they share uncommon attributes). They apply the algorithms to a relational data set of employment records to identify tightly knit groups of people who are at high risk for fraud. This work can be classified as a relational knowledge discovery scheme [11] that utilizes the relationship between individuals and their attributes to infer patterns and make predictions.

This technique is widely used in the natural sciences; for example, it is used to determine family structure in animal groups based on animals sighted together [14].

Zheleva and Getoor [24] explored inferring user attributes in a social network, using a number of different user profile elements. When attempting to infer a user's attributes from the social network alone, they only consider the revealed attributes of the user's friends whereas we also consider the attributes of users who are not directly connected to the user in question. Moreover, the most successful approach they find relies on having additional information about users (such as group memberships). As a result, their method is complementary to our approach (which requires information about the social network and partial user attributes).

Researchers have also examined the opposite problem: using user attributes to predict social links. For example, Adamic and Adar [1] used similarity in the text and links on users' web pages to predict the likelihood that users are friends. This work is orthogonal to ours, as its goal is to infer the social network graph, whereas we assume that this graph is known.

6. CONCLUSION

In this paper, we examined the question: given attributes for some fraction of the users in an online social network, can we infer the attributes of the remaining users? Using fine-grained data taken from two large online social networks, we found that users are often friends with others who share their attributes. Moreover, we found that communities form in the network around users who share certain attributes. These two observations lead to a natural approach for inferring user attributes, namely, to leverage automatic community detection in order to infer attributes.

However, we found that existing approaches were not able to detect communities centered around common attributes in all cases. Thus, inspired by previous work on community detection, we proposed a new approach for detecting communities that is able to detect communities for multiple attributes in our data set. In fact, we found that, with as little as 20% of the users providing attributes, we could often infer the attributes for the remaining users with over 80% accuracy. We make our algorithm implementation and data sets available to the community.

Our work has a number of implications and uses. For example, many of the popular online social networks could directly apply our algorithm in order to detect certain attributes for users who do not provide them. This would enhance the user experience on the sites, as the attributes provided are often used for guiding search results, for suggesting users who may benefit from interaction, and for grouping users. Moreover, it could also be used to reduce the current burden on users imposed by manual data entry.

However, our findings also raise interesting questions about the nature of privacy in online social networks. In particular, almost all privacy mechanisms available to users today are based on access control: users can specify which other users are able to view the content or information they upload. Our results show, however, that even information that is not provided by users can sometimes be inferred from the user's location in the network. Thus, it is not sufficient to ensure privacy by making attributes private, instead, both attributes and the list of a user's friends must be marked private to ensure that a user's attributes cannot be inferred.

7. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] R. Andersen and K. J. Lang. Communities from seed sets. In *Proc. WWW'06*, Edinburgh, Scotland, May 2006.

[3] J. P. Bagrow. Evaluating local community methods in networks. *J. Stat. Mech.*, 2008(5), 2008.

[4] A. Clauset. Finding local community structure in networks. *Physical Review E*, 72, 2005.

[5] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), 2004.

[6] O. Şimşek and D. Jensen. Navigating networks by using homophily and degree. *PNAS*, 105(35):12758–12762, September 2008.

[7] Facebook. `http://www.facebook.com`.

[8] A. T. Fiore and J. S. Donath. Homophily in online dating: when do you like someone like yourself? In *Proc. CHI'05*, Portland, USA, 2005.

[9] A. L. N. Fred and A. K. Jain. Robust data clustering. In *Proc. CVPR'03*, pages 128–133, June 2003.

[10] L. Friedland and D. Jensen. Finding tribes: identifying close-knit individuals from employment patterns. In *Proc. KDD'07*, San Jose, California, USA, Aug 2007.

[11] D. Jensen and J. Neville. Data mining in social networks. *In Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 287–302, 2003.

[12] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, May 2004.

[13] F. Luo, J. Z. Wang, and E. Promislow. Exploring local community structures in large networks. *Web Intelligent and Agent Systems*, 6(4):387–400, 2008.

[14] D. Lusseau and M. E. J. Newman. Identifying the role that individual animals play in their social network. *Proc. R. Soc. London B*, 271:S477, 2004.

[15] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. IMC'07*, San Diego, CA, October 2007.

[16] M. E. J. Newman. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, August 2001.

[17] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), 2004.

[18] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[19] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *PNAS*, 101(9):2658–2663, March 2004.

[20] Rice Culture. `http://www.professor.rice.edu/professor/Rice_Culture.asp?SnID=165470151%4`.

[21] Rice University Alumni Directory. `https://online.alumni.rice.edu/directory/detailsearch.asp`.

[22] Rice University Student Directory. `http://www.rice.edu/search/query.php?advanced=1&tab=people`.

[23] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *Proc. ICCT'03*, Dordrecht, 2003.

[24] E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proc. WWW'09*, Madrid, Spain, May 2009.