



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<sup>2</sup>, and major(s) or department.

<sup>2</sup>Rice University has nine residential colleges, to which incoming undergraduate students are randomly assigned. The colleges serve as dormitories, cafeterias, and social circles;

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.<sup>3</sup> 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.<sup>4</sup> 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.<sup>5</sup> 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.

students stay at the same college during their entire undergraduate tenure.

<sup>3</sup>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.

<sup>4</sup>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.

<sup>5</sup>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*<sup>6</sup>, ranges from 0 to  $\infty$  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

<sup>6</sup>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  $\text{Tr } \mathbf{e} = \sum_i e_{ii}$  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.<sup>7</sup> 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

<sup>7</sup>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

$$\frac{-2 \sum_i \sum_j \mathbf{x}_{ij} \log\left(\frac{\mathbf{x}_{ij} N}{\mathbf{x}_i \cdot \mathbf{x}_j}\right)}{\sum_i \mathbf{x}_i \cdot \log\left(\frac{\mathbf{x}_i}{N}\right) + \sum_j X_j \log\left(\frac{X_j}{N}\right)} \quad (4)$$

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  $\mathbf{x}_j$  denote the sum over column  $i$  and row  $j$  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









