Project Number: 46-SMS-0601

ANALYSIS OF SOCIAL NETWORKS USING FACEBOOK

An Interactive Qualifying Project Report
submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Bachelor of Science

by

Christian Roy

Jonathan Adler Christian Roy

Date: March 1, 2007

Stanley M Selkow

Professor Stanley M. Selkow, Major Advisor

1. social networks
2. graph theory

DJ Dougherty

Professor Daniel J. Dougherty, Co-Advisor

Abstract

College is an important time of social growth in a person's life. Often, connections formed in college will have a large impact later in life. As such, it is important to understand the social dynamics on a college campus. These dynamics can be viewed as Social Networks, an increasingly important branch of the social sciences. In this project we intend to apply mathematical analysis to a Social Network in order to study the implications they have on everyday life.

Contents

List of Figures

List of Tables

1 Introduction

Mathematics is often used to model everyday phenomena. In our project, we aim to apply the field of mathematical science known as graph theory to a social network. This will allow us to utilize the vast amount of research done for graph theory in a social context.

1.1 Social Networks

A *social network* is a collection of people or organizations and the ties between them. These ties may be anything from riding on the same subway to having a close personal friendship. One of the primary uses of social networks is as a model for analysis, as it can provide a easy means to visualize and associate related people and groups[13]. Social networks can help one many different situations and relationships that occur between people. As an example, using social networks the importance of key people in a social environment can be easily recognized.

The Internet is one of the fastest growing areas in which social networks may be viewed. In this case, social networking web sites are designed to make it simple for users to create their own social networks via related interest or shared media. Examples of these sites range from news based such as Digg.com and Reddit.com to music based like last.fm. As the Internet moves to a more user-driven paradigm, social networks are becoming a central theme.

1.2 Graph Theory

In mathematics and computer science, a graph G is formally defined as $G = (V, E)$ where V is a set of vertices (or nodes) and E is a set of ordered pairs of nodes. The set E is defined as the set of *edges* where an edge represents a connection between two vertices[7]. Graph theory is the field of mathematics based around these graphs. Within the field of graph theory there are many categories of graphs. The simplest type, called a *simple graph*, is one in which for any two nodes, there can be at most edge between them. In a simple graph a node may not have an edge to itself. Such edges are called *loops*. A *multigraph* is one in which there may be multiple edges between any given pair of nodes, but the graph still may not contain loops. In a

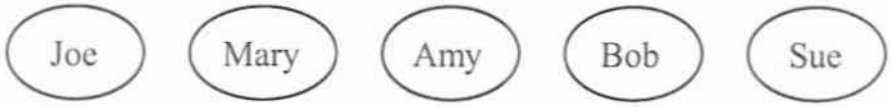


Figure 1: A graph with no edges

pseudograph, there may be multiple edges between a pair of nodes and the graph may contain loops.

There are also ways to include more data about edges in the graph. One of these is through a *directed graph* or *digraph*, in which each edge includes a direction. That is to say that node A may connect to node B, but node B may not connect to node A. In addition, there are also *weighted graphs*, in which each edge is associated with a numerical value. There also exist *weighted digraphs*, which include both a direction and a weight for every edge.

Another piece of information we can study about a graph is the existence of *cliques*. A clique is defined as a complete subgraph. A complete subgraph is a graph such that all the vertices have an edge to each other vertex. If we have six vertices and all have edges connecting to each other, then the six vertices would be a clique of order six.

As can be seen, the field of graph theory can be quite diverse. It is however, a powerful model for visualizing structures and has applications in many different areas. More information on graph theory and its applications may be found in Introductory Graph Theory[8].

1.3 Graph Theory and Social Networks

Graph theory provides an interesting method in which to model social networks. If we take each person or organization and make them a node, then we can represent the ties as edges. These edges may furthermore be given weights to represent strength or frequency of the connections, forming a weighted graph. For this report, we use the term *friendship* when speaking about the edge between the two nodes and the term *friend* when we speak about the person to whom it is connecting. As an example, we may consider a small group of people Joe, Mary, Amy, Bob, and Sam. We may then add a series of friendships, creating a simple graph. In the following example, Mary is friends with Amy and Joe, Joe is friends with Mary, Amy, and Sam, and Bob is friends with Sam. This is a very simple model, but it can easily

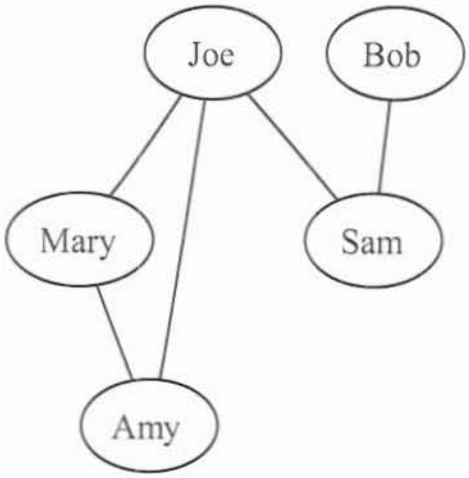


Figure 2: Adding edges

show who is the most popular in this group, merely by counting the number of edges at each node. This calculation is the *degree* of a node. In this case, Joe would be the most popular as he is friends with three of the other four. However, we can also represent a more complicated social environment by dropping the restriction that friendships are symmetric. In this case, a person may feel he is friends with someone who does not share that belief. In order to model this, we can use a digraph.

If we use this new model, we can represent a more complex social situation. In this case, we may add that while the previously defined friendships are still symmetric, Bob also considers Amy a friend, resulting in the following graph. With this graph, we can again see who is the most popular by checking the number of edges ending at the vertex. With this graph, however, we can also easily see who may consider themselves better connected then they actually are. These will be the people with a larger outward degree than inward.

However, we can also increase the complexity of this graph even more by using a weighted digraph. For each edge, we can assign a weight based on how much the person the edge originates from likes the person it is directed to. This permits an even more in-depth form of analysis, and can differentiate between those who are barely considered friends and those who are very close. Even more, with this model we may obtain a more accurate representation of

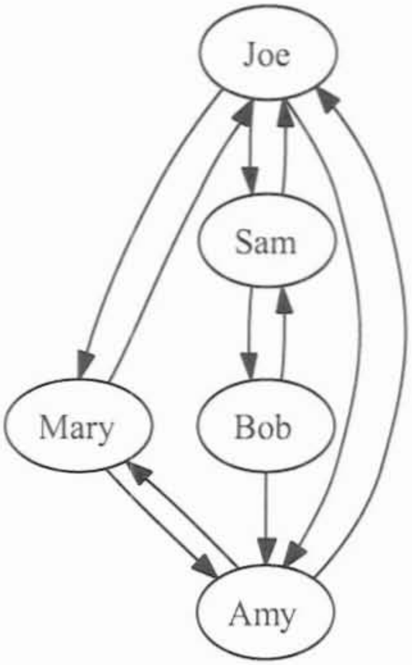


Figure 3: Adding direction

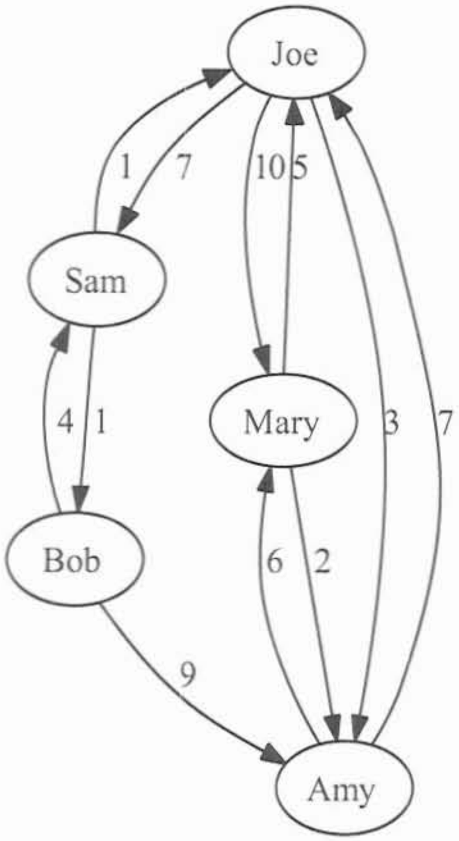


Figure 4: Adding weights

popularity. Instead of merely counting edges, we may add the weights of the edges directed towards a node as a popularity score and rank accordingly. While this is a rather trivial example, more complex analyses may be done. Data such as this can be used for anything from deciding on a guest list to determining sitting order in a movie theater. Even more analysis may be done, depending on what we use to determine edges and their weights.

1.4 Intended Analyses

Since we can model social networks as graphs, we can also use any graph theory algorithms in our analysis of them. Moreover, we may be able to take an algorithm from graph theory and look for its meaning on a social network. For most of our analysis, our vertices will be individual people, and the edges will be friendships. In order to simplify the model, we will assume that if person A is friends with person B than person B is also friends with A. Moreover, we will assume that there are not 'degrees' of friendship, that is to say we will be treating all friendships as having equal weight. In this way, we will create a simple non-directed, non-weighted graph of a social network to perform analysis on.

Once we have this graph, we can obtain a large amount of useful information from it. By merely investigating it visually we can find where there are groups of friendships by seeing a large cluster of edges going between a series of vertices. We can also look for large subsets of vertices without connecting edges showing that groups of people do not know each other. A single vertex of high degree means that the person involved has many friends.

We can also use the graph to calculate how efficient a person would be at sending a message to all the people represented in the graph. Consider two people, Carl and David, who are both running for the position of president of the campus. Carl has 100 friends, while David has only 50. Although it seems that Carl has the advantage, since he knows more people who he could promote his candidacy to, that is not necessarily the case. If each one of Carl's friends has only 3 friends besides Carl, while each one of David's friends has 20 friends besides David and each other, a situation can be constructed in which David is better suited to promote his candidacy. As an example, if we consider the possibility that friends of the people running will tell their friends whom to vote for, then Carl would have 300 votes while David would have 1,000. So despite Carl knowing more people, the people he knows are not as well connected as David's friends and therefore it is possible he cannot send out a message as well as David. An example graph in which this situation can occur can be seen in Figure 1.5. When put into graphical form, it is important to not only consider the degree of a vertex, but also where the edges connect to. The number of vertices that are distance two away, where distance is defined as the minimum number of edges one must travel to connect two vertices, can be as important as the degree of the vertex.

In addition to studying the edges, we may also look at cliques. Socially, a clique would represent a group of people who all know each other. We hypothesize that large cliques would be most likely found in clubs or organizations on campus. Social environments such as clubs or organizations allow for many people to meet and interact with each other simultaneously. If these interactions become friendships then a clique would be formed. It

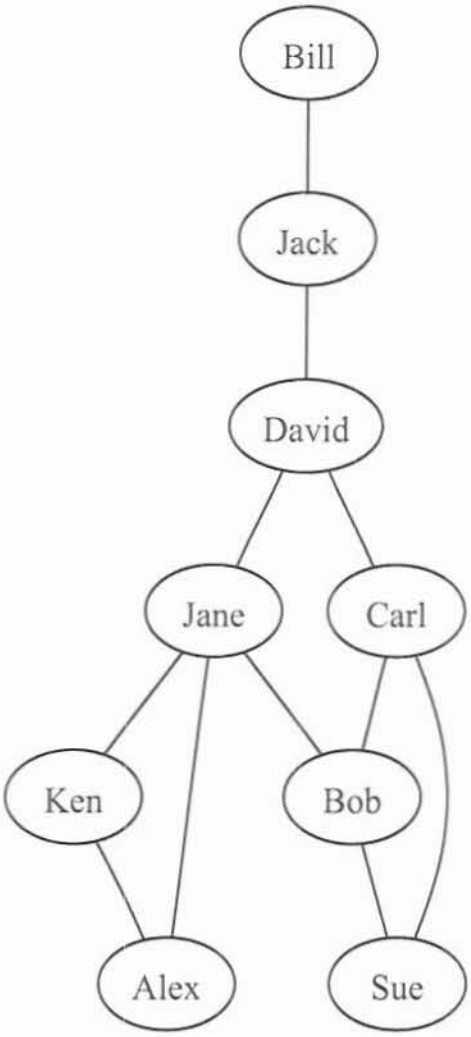


Figure 5: An Example of Distance In A Graph

would be much more difficult for a clique to form with people who don't all have a reason to interact, since it is difficult to make a connection without interaction.

In addition to these, we would also like to study the differences between the social networks at different schools. Unfortunately, due to problems acquiring the data required to make graphs at other colleges, we were only able to analyze the social network at Worcester Polytechnic Institute.

2 Gathering Data

In our research we considered several possible sources of data, as well as different means of performing the analysis. In this section we will discuss the different choices we made and why we made them.

2.1 Social Networking Web Sites

In order to analyze social networks on college campuses, it is necessary to have access to information about these networks. Therefore, we must have a means of finding friendships in order to form a graph of them. Social networking web sites have been increasing in popularity and as such we were able to choose from one of them as a data source. The three we primarily looked into were MySpace[4], Facebook[2], and Livejournal[3]. Data we considered while making our choice included both the size of the user base as well as the site's Alexa[1] ranking. An Alexa ranking is used to measure the amount of traffic a site receives over a given period of time. Alexa works by installing a client on volunteers' computers and then tracking what web sites they go to anonymously. The program then makes a list of how frequently different web sites are visited. The Alexa ranking only represents the amount of traffic a site receives, not necessarily how many different people actually use that site. In other words, if a site has ten very active users, it will have a better Alexa ranking than a web site that has twenty inactive users. In terms of Alexa rankings, an ordered list is created based on traffic, with lower elements in the list having a better ranking. As such, a web site with ranking five is better than one with ranking eleven. The site itself may be visited at http://www.alexa.com for more information.

2.1.1 Myspace

MySpace is a social networking web site created in 2003. Its aim is to allow users to share both information and multimedia with other users. It considers friendship to be a symmetric relationship and supports an easy interface for listing other users as friends. With a user base of over 100 million people at the time of this writing, it has an Alexa ranking of 6, making it one of the most popular sites on the Internet. Unfortunately, while MySpace has support for storing information about what college a user attends, there is no checking on whether they actually attended that college or not. Moreover, there is no advantage or need to list which college a user attends on their MySpace account. As such there is a certain amount of noise in the data that would hinder the work done on this report.

2.1.2 Facebook

Facebook, like MySpace, was specifically created as a social networking web site. It was created in 2004 and currently has a user base of over 10 million people. It also has an Alexa ranking of 54, making it less popular than MySpace in both traffic and users.

Facebook's initial goal was to create a place where college students could connect with each other on the Internet. In order to join, it used to be required that one had a valid e-mail address from a Facebook recognized college. While it has since expanded to allow the general public to join, there are still advantages to declaring a college, such as allowing further access to information about other students at that college. Since one needs a valid e-mail address from that college to be listed as a member on Facebook, it becomes harder to trick the system, reducing noise.

Facebook has also recently provided Application Programming Interface (API) access to some of their stored data, further simplifying the prospect of obtaining it. Their API allows us to get information about a given user such as who their friends are and what clubs they are in.

2.1.3 Livejournal

Livejournal was created in 1999 and has an Alexa ranking of 63, which means it receives the least traffic out of any site we considered. Furthermore, the original aim behind Livejournal was not social networking as such; rather it was a means for people to keep online journals. All social networking, as a result, is done solely through reading other peoples journals. While it is possible for one to list friends in order to simplify this task, it is almost a secondary consideration.

With a user base of over 11 million accounts (1.8 million active), it has a similar number of users as Facebook. However, these users are much less active than the users of Facebook, resulting in both a lower Alexa ranking and most likely a lower degree of accuracy, since if the site isn't used much people are not updating to current information. Similarly to MySpace, Livejournal also allows you to join groups based on your college. Unfortunately, there is also no check in place to verify that you do actually attend that college.

2.1.4 Conclusion

We concluded that Facebook is the ideal data source for this project. While the number of users on MySpace is substantially greater, since there is no verification on the data the accuracy would be low. For similar reasons, Livejournal introduces too much error to use as well. Since Livejournal has roughly the same number of users as Facebook, their is no advantage to using it over Facebook. Facebook, in addition to having the cleanest data, also provides the easiest access to that data via its API, making it a clear choice.

2.2 The Data Collection Process

In order to gain data about friendships from Facebook, we first needed to obtain the unique Facebook ID for each individual we wanted data on. At first glance, this appeared to be a trivial exercise since on Facebook's main site, the User IDs are numeric and sequential. However, upon further research it became apparent that the User IDs used by the API were different from those on their primary site. These new IDs were non-numeric and non-sequential, making it impossible to obtain without some input from a user. Since we were incapable of determining a user's ID without some input from either that user or his or her friend. Due to these difficulties in obtaining data, we shifted our focus to study the WPI campus specifically.

In order to gain the IDs of different students, we set up a table in the middle of the WPI Campus Center, a location many students pass through daily. Throughout the course of a week, we asked people to log in to assist us in our project. Unfortunately, the number of people willing to help was relatively small compared to the size of the campus. With that said, we were also able to get the Facebook IDs of each of our volunteer's friends as well. Because we were only studying WPI, we only retrieved Facebook IDs of WPI students. As a result, event though only approximately 120 people were willing to assist us, this resulted in us obtaining 2259 unique Facebook IDs, over half of all WPI students on Facebook. We consider this to be a sufficiently sized subset of the campus, and as such moved on with our research and analysis.

2.2.1 Getting Friendships

Having obtained a listing of Facebook user IDs, we still had the challenge of determining which people were friends. Fortunately, Facebook has an API call that takes two same-sized lists of user IDs. This call returns a list of the same size back. For each element of the list returned, if the two user IDs at that index in the lists sent are friends it has a true value, otherwise it is false. Since we need to check all possible friendships, that means that we need to create a series of lists such that all possible friendships are represented between them. Since friendship is a relationship between two different people, we can also exclude the possibility of someone being friends with themselves. Upon first examination, it appears that this would result in an $(n-1)^2$ calculation, where n is the number of people. However since friendships must be symmetric, we only need to check under half that. In fact, since for the first person we only need to check friendships with the remaining $n-1$ people and for the second person we only need to check with the remaining $n-2$ and likewise throughout the rest of the people, the number of calculations required can be represented by the summation $\sum_{i=1}^{n-1} i$, a calculation we can reduce to $\frac{n*(n-1)}{2}$. Since we have 2259 people, this calculation reduces to $\frac{2259*2258}{2} = \frac{5100822}{2} = 2550411$.

This means that for the number of people we obtained IDs for, there are over two million possible friendships. As we needed to check all possible friendships, it became clear that the task should be broken into manageable chunks. A large number of factors went into deciding how many friendships we should check at each time, but the primary one was the number of calls to Facebook's API it would take, as we were only allowed 100,000 API calls daily. Our final decision was to check three thousand friendships at a time. Since we were getting this data in chunks, we needed some way to store each chunk. In this case, we used Python's serialization features to save each consecutive list to a separate file. This made recombining the lists after all the data was gathered a trivial task.

Testing all possible friendships took over two hours, largely due to transit time on the Internet. In the end, we found that of the 2,550,411 possible friendships, only 84,293 of them actually existed. This means that less than 4 percent of all possible friendships exist, a number that we found surprisingly low.

2.3 User Privacy

When dealing with data about people and their social ties, privacy is an important issue. Some people would be uncomfortable with being specified directly via analysis. Even more, since social groups and connectivity can be an important link to personal identity they are often taken very seriously. Research showing that people may be disconnected or even friendless may offend or threaten some people, and as such it is important to avoid singling out any specific person.

Because privacy is such a concern, we took it very seriously throughout our research. As such, we put forth every effort to avoid specifying any person throughout our analysis. While gathering the data, we assigned each person a unique number and referenced them through that instead of by name in our algorithms. After assigning these unique numbers, we deleted the name data from our records entirely.

Since we are studying the whole of a social network instead of specific individuals in one we performed all analysis in aggregate. Even though we also discuss club impact and analyze a social network of clubs on campus, a club is still separated from the people that make it up.

An additional measure of privacy was gained from using Facebook.com as our data source. Facebook allows its users to customize greatly the amount of access users have to their information. Even more importantly, it offers specific options to users regarding changing the access level someone using the API has to their profile. Additionally, the API blocks access to certain information about users regardless of user preference in this regard.

2.4 Language Choices For Obtaining Data

Facebook's API uses a simple Representational State Transfer (REST) interface. A REST interface is simply a way of making remote function calls over Hypertext Transfer Protocol (HTTP), without including an additional messaging layer on top. In Facebook's API, requests are sent in an Extensible Markup Language (XML) format to the server. In order to make these requests easier to use, several people have created client libraries in a variety of languages and have offered them for use to the general public. In order to save time and not recreate already completed work, we decided to use one of these libraries. This limited our choices to either C#, C++, Cocoa, Java, Perl, PHP, Python, Ruby, or Visual Basic.NET.

2.4.1 Scripting Versus Compiled

Our first major decision was whether to go with a scripting language or a compiled language. In this regard, Cocoa, C#, C++, Visual Basic.NET, and Java were considered compiled, with the rest being scripting. In this case, we decided to go with a scripting language because they frequently offer a faster development cycle (via eliminating the compilation aspect) at the expense of runtime. Since network speed and file I/O were likely to be the limiting factors on this stage of the project, we decided that execution time was less important than rapid prototyping. Having decided on a scripting language, our options were thus cut down to Perl, PHP, Python, or Ruby.

2.4.2 Conclusion

In the end, we decided to go with Python because of its simple, expressive style. The API library provided was both intuitive and powerful, and the Python language supports operations such as list slicing and iteration in a simple way. This enabled us to focus on algorithms for obtaining the data as opposed to wrestling with how to accomplish a given task in the language. Moreover, Python's clean object oriented interface and dynamic typing helped our code be even more expressive and intuitive than it would have otherwise.

3 Errors in Modeling the Social Network of WPI

Our model of WPI, although very useful, has some inaccuracies due to the method of obtaining the data. First, there are inaccuracies in the friendships we found because of using Facebook. Oftentimes people become Facebook friends merely because they noticed they had shared interests or were in the same class, despite having never spoken in person. Also, it is possible for two people to fall out of friendship in real life while still claiming to be friends on Facebook. While it is possible to remove a friendship on Facebook, this is seen as an impolite act and as such rarely occurs. Another problem with the friendships is that two people could talk to each other all the time and consider themselves friends however not be Facebook friends.

Another source of error lies in our method of obtaining the data from Facebook. Because of the method of obtaining data, namely manually asking people to login for us, we have a bias towards people we know. This is because since they were our friends they were more willing to give us their time to log into Facebook for us. So because of this the graph has a bias towards having people who are closely connected to the test givers in the graph while leaving out people who have a high distance from the surveyors. Additionally, there is a problem if there was a person who never goes into the campus center nor do any of her friends she would not be in our graph.

As well as these sources of error, there is some potential error due to the changing of information on Facebook. When we did our survey, we originally obtained unique Facebook IDs for 2259 individual people. However, since the course of the survey ran throughout an entire week, when we then used 25 of these IDs to check who they were friends with, we got nothing back. While this could be the incredibly unlikely situation that none of these 25 people have a single friend and all happened to take our survey, it is substantially more likely that either they have that data blocked to the API or that their ID had changed at some point throughout the week. As we were not able to use these IDs to gain information, they would just end as disconnected vertices on our graph and as such we discarded them from our dataset. This leaves us with 2234 remaining vertices, representing the final dataset we performed analysis on.

An interesting trait of our graph is that it shows the social network of WPI for November 16th, 2006 at one specific moment in time. Although this is useful, social networks are constantly changing as new friendships are formed and old friendships are broken. Our data will quickly become outdated as the social network of WPI changes. Although having said that our data is still useful for analyzing the social network on that day.

Table 1: How many people can be reached

	Average Number of People	Standard Deviation	Max	Min
Distance 1	75.46374217	59.70507289	491	1
Distance 2	1399.620412	394.7802634	1954	40
Distance 3	750.8415398	427.1609418	1931	20
Distance 4	7.071620412	37.73439681	1009	0
Distance 5	0.002685765	0.059794713	2	0

4 Distances Between Nodes

With our data, we were able to create a 2234 by 2234 matrix of adjacencies. This matrix had a column for each node, as well as a node for each row. So the value A(i,j) represented the edge between nodes i and j, with a 1 meaning that an edge existed between the two, while a 0 represented no edge. A matrix of adjacencies was chosen for this purpose because it allowed for a simple means of calculating distance between nodes, as well as providing a way to check if two vertices were connected. Squaring this matrix, by which we mean to multiply the matrix by itself, gives us the number of paths of length 2 between i and j. Cubing this matrix showed us the number of paths of length 3 between i and j, and so on. Using these exponents of the original matrix allowed us to easily find the distance between any two vertices. Originally, Dijkstra's algorithm[9] was considered in order to find the path length between two points. However, the algorithm is more complex than simple matrix multiplication and the time gained was not worth the additional complexity. Socially, this minimum path would represent the quickest way for two people to get in contact with each other by only talking to their friends. For example, if Al was trying to tell people about his exciting watermelon eating event on campus and he told everyone he knew would the message get to Bob, or would everyone he knew have to then spread the message before it got to Bob? One can easily see how this chain may continue across friends until the message was delivered.

These charts shows the results of our distance matrix. The average person tends to have most people within two steps, although there is a large deviation within our test group. There are very few pairs of vertices with distance 5, which is the largest distance we found. No two people were disconnected from each other, although the possibility for a bias in the sample we took could account for this.

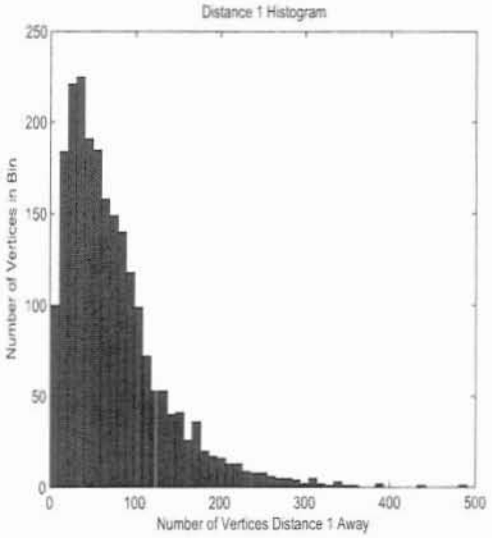


Figure 6: Number of Vertices Distance 1 Away

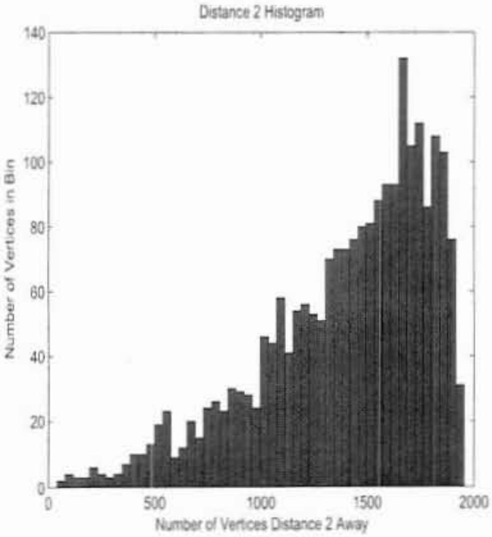


Figure 7: Number of Vertices Distance 2 Away

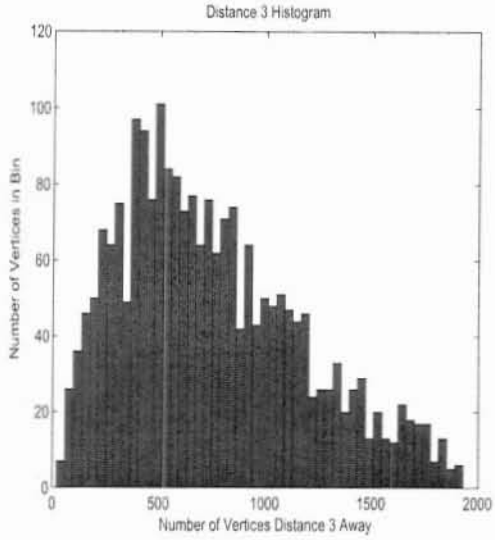


Figure 8: Number of Vertices Distance 3 Away

A question we proposed was how important is it that a person knows the right people to be connected to the campus. While a more complete analysis is shown later in the section on centrality, one can see that it does in fact matter the degrees of the vertices a vertex is connected to in terms of message passing. As an example, with this graph it was easily seen that two people with the same degree may be connected to a vastly different number of people via a path of distance 2.

This image shows a scatter plot where each vertex represents a person in our database. The x-axis is the number of friends that person has, while the y-axis is the number of people who are distance two away from that person. There is a clear cluster here, however the cluster is spread out significantly. There are people who have approximately 50 friends and 700 people of distance two and there are people who have around 50 friends and 1700 people of distance two. Although there is a visible center to this cluster, the overall spread of it signifies that who you know matters as well as how many people you know.

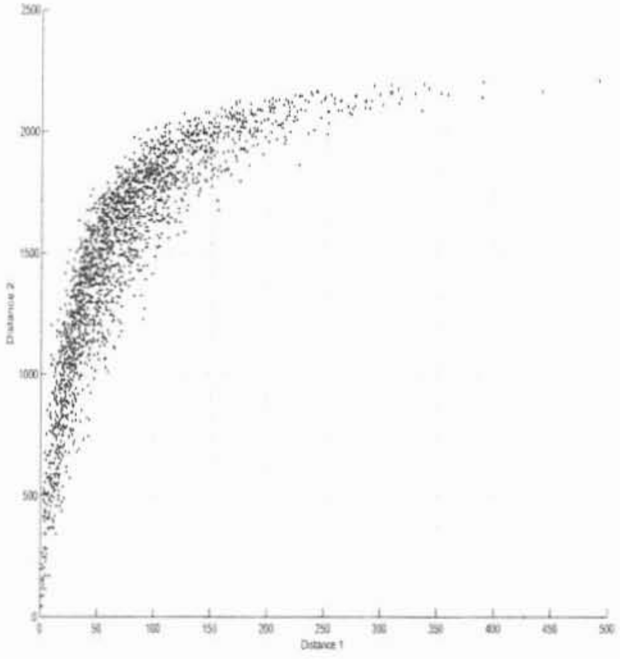


Figure 9: Scatter plot of Distances 1 and 2 For Any Vertex

5 Centrality

The concept of centrality of a node is based on the idea of someone being at the center of a social network. This person is "well connected" in that they have edges connecting to nodes all over the graph. Centrality is merely the measure of how well connected a person in the graph is, where a person with high centrality is either connected to everyone, or is only a few edges away from a large number of people. A person with a high centrality would have an easier time sending a message across the rest of the graph since he is well connected to all of them, while a person with low centrality would have more difficulty.

There are many different methods of calculating centrality in graph theory. One major branch of the methods is known as degree based centrality, where the degree of a vertex is used to calculate how central a vertex is in a graph[6]. A person with many edges has a high centrality since that person is connected to many different people in the graph. Socially, the more friends a person has the more central in the graph this person is, according to this model. In our data, the range of number of friends each person had was from 1 to 491.

However this method of centrality does not take many factors into account. A person named Al can have many friends, however if those friends are not adjacent to anyone else in the graph then Al does not have contact with a large portion of the graph. Likewise, if a person named Bob has only five friends but those five friends are highly central, then Bob is well connected in the graph in that he can use his five friends to communicate with the rest of the graph.

A better method of calculating centrality is to use a measure of distances between vertices. It is easy to calculate the minimum length of a path between two vertices by doing a calculation involving raising the adjacency matrix to a positive integer power. We can define centrality as the sum of the distances from a vertex to all other vertices. It is important to note that this calculation of centrality only works for connected graphs. If a graph has nodes that are not connected to each other via a path of any length, then it would be impossible to calculate the distance between them (since there is no path, conventionally this distance is considered to be ∞).

$$c_i = \sum_{j=1}^{n} r_{i,j}$$

In this equation c_i is defined as the centrality of vertex i. $r_{i,j}$ is defined as the distance between vertices i and j. A lower value for centrality means that the vertex is better connected, and thus more central in the graph. This method of centrality is more in line with our definition of centrality because it measures how easy it is to send a message across the entire graph, with a lower score meaning it takes less transitions across vertices[12]

Another method based on this type of calculation was published by Bonacich in 1972[5]. He stated that these sums needed a weighting factor based on the centrality of the person our vertex is connecting to. So to be far away from a person who is well connected is much worse than being far away from a person who has only a few connections. The new method of centrality calculation is as follows:

$$c_i^1 = \sum_{j=1}^{n} r_{i,j} c_j^0$$

For this equation, c_j^0 represents the centrality of vertex j calculated by the previous equation. c_i^1 is the new centrality according to this calculation. The centrality then is defined by an initial calculation of centrality using the formula proposed by Sabidussi. When compared to the data from the first method the two centrality calculations had similar results. We were curious as to what would happen if we did a third calculation of centrality in which we repeatedly used Bonacich's method on a centrality calculation. This method would work iteratively, starting with the initial case defined by Sabidussi. To keep the results from growing exponentially, after each calculation we would divide all the centralities of that iteration by the maximum value achieved during that iteration. Thus all of the centralities would be mapped onto the interval [0,1] which allows is to compare centralities from different iterations.

$$c_i^k = \sum_{j=1}^{n} r_{i,j} c_j^{k-1} / max_m(c_m^{k-1})$$

$max_m(c_m^{k-1})$ is defined of the maximum value of c_m^{k-1} for all m. The result of using this iterative method was that the centralities quickly converged to a final result. By the tenth iteration, there were no changes to the centrality values we were calculating. Using a least squares calculation, the combined distances of centrality values from the original centrality calculation to the tenth iteration calculation was only .30 which is quite surprising considering

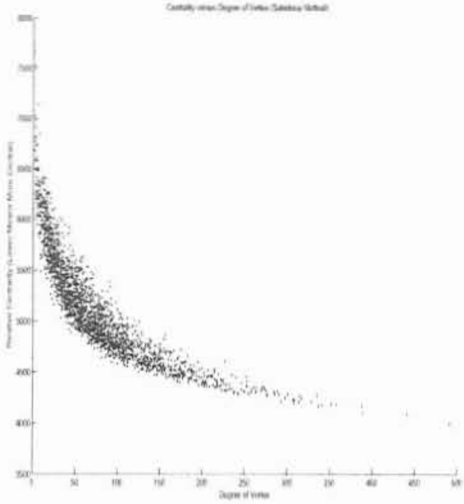


Figure 10: Centrality versus Degree of Vertex (Sabidussi Method)

that each vertex has a centrality of around .7. So there was not even much change in the iterative method compared to the original method proposed by Sabidussi, which calls to question the necessity of the iteration of Bonacich's method.

When we used these methods of calculating centrality on our graph we got interesting results. Using the model proposed by Sabidussi, we calculated the centrality of each vertex on our graph, and got centrality scores ranging from 3995 to 7664.

These scatter plots illustrate the fact that degree of a vertex is not directly correlated to centrality of a vertex. The thick width of the cluster of vertices with degrees between 25 and 150 is because vertices that might have the same degree do not necessarily have the same centrality.

We were also able to apply these measures of centrality to groups of vertices instead of individual vertices. Using the clubs and organizations of WPI that we obtained from our data gathering, we could test these groups to see which ones were more central to the graph. To study the measure for each club, we found the minimum distance between each student on campus to any member of the club. This is the same as taking all the members of the club, combining them into one vertex (with edges connecting to all of the

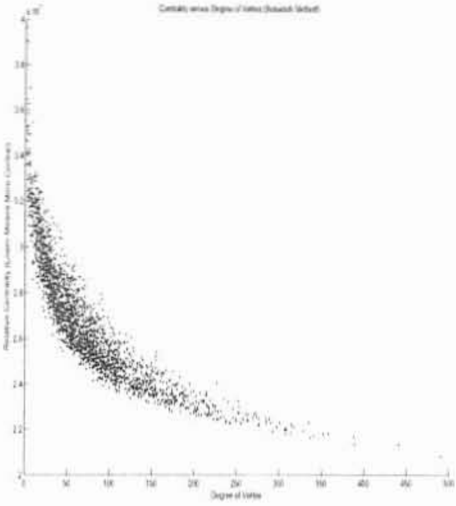


Figure 11: Centrality versus Degree of Vertex (Bonacich Method)

vertices that members of the club had edges to), and applying the centrality calculations. With these calculations we were able to rank each of the clubs' centrality relative to the graph of the campus such that a higher rank meant that the club had more central members. Additional analysis on clubs and centrality is found in our next section.

6 Using Randomly Generated Social Network

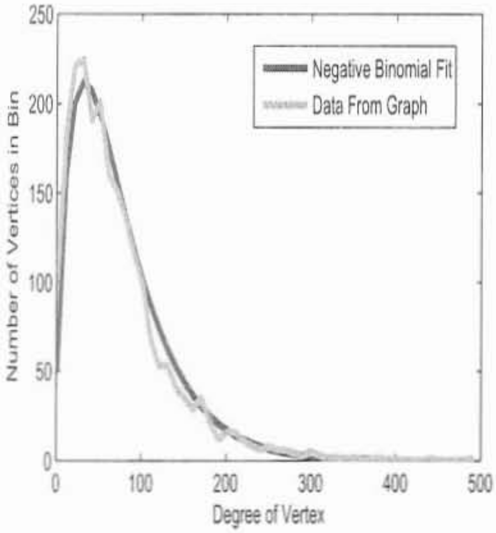
One of the questions we wished to investigate with this research was if it would be possible to generate an artificial graph that had similar traits to the social network we created from real world data. By doing so we would be able to find what characteristics defined the graph.

Our first method of recreating the graph was rather crude. We created a graph with 2234 vertices, then inserted 84293 edges in the graph by randomly choosing the starting and ending vertices. Although this graph had the same average number of edges per person in it (at 75.5 edges coming out of each vertex), the standard deviation of number of edges connecting to a vertex was 8.55 which is nowhere near the standard deviation for the degree of a vertex in our graph. This method does not make much sense from a social networking standpoint because it assumes that all of the people in the network have the same social setting and abilities. That is to say, any two people are just as likely to be friends as anyone else. In a social network however there are people who are more social, less social, and some people will more have more edges than others have higher degrees. Our first method did not model this at all.

Our second method was to study the average degree of the vertices in the graph and see if we could find a method of randomly generating a number of edges for a vertex and then connecting those edges to the rest of the graph. If we found a function to give us a number of edges then we would have more accurately modeled the amount of friends a person had, even if we had not figured out the method in which the friends were chosen it would still be an improvement to the first model.

Table 2: Comparison of Random Graphs to Our Data

	Average Degree	Average Centrality (Bonacich)	Average Centrality (Sabidussi)
Random Graph 1	75.55	4577	$2.07 * 10^7$
Random Graph 2	69.55	4911	$2.43 * 10^7$
Facebook	75.46	5155	$2.68 * 10^7$



Upon inspection of the histogram created by the degree of vertices and discussion with a professor of statistics, it seemed that the histogram closely resembled a negative binomial distribution. A negative binomial distribution is used to describe the number of trials that will occur before a predetermined number of successes is achieved by a series of independent trails[11]. At the moment of this writing we are not sure precisely why the negative binomial distribution models the histogram so well, but regardless it had a sound fit, as was seen when we did a least squares analysis on the data.

With these two types of randomly generated graphs we were able to run tests of centrality to compare with our data from Facebook.

The first random graph had an average centrality considerably lower than the data from Facebook. This was the graph created by arbitrarily choosing starting and ending nodes for our edges until we had placed the same number of edges that were in the graph given by the Facebook data. So the fact that

Table 3: Comparison of Random Graphs to Our Data

	STD of Degree	STD of Centrality	STD of Centrality
		(Bonacich)	(Sabidussi)
Random Graph 1	8.56	59.46	$0.271 * 10^7$
Random Graph 2	48.36	431.28	$2.08 * 10^7$
Facebook	59.71	502.70	$2.57 * 10^7$

the centrality is lower is interesting because it means that there are methods in making vertices more centralized in a graph then by whatever method is used to create friendships in a social network. A social network is not optimized for average centrality. The first random graph also had a lower standard deviation of centrality than the Facebook data. So the vertices all had a very close centrality (with an order of magnitude less of standard deviation than the Facebook data). There was an interesting discrepancy in the vertex of minimum centrality of the randomly generated graph. The most central point in the random graph using the Bonacich method had a value of 4413. The most central point in the Facebook data had a value of 3995 using the Bonacich method. So using the Bonacich method it appears that if we wanted to make a graph with a single vertex that was as centralized as possible then the Facebook data would be better. However when we check for centrality using the Sabidussi method we find that the random graph had a vertex of lower minimum centrality than the Facebook data did, which contradicts our results using the Bonacich method. This is perhaps an indication that the two methods of calculating centrality can lead to discrepancies since they do not measure quite the same thing.

The second random graph was more similar to the data from Facebook in many ways. Although it's average vertex degree was lower than the Facebook data, the standard deviation of degree was only 10 lower than the Facebook data (As opposed to the 50 points off of the first random graph). On the matter of centrality, the second random graph had a average centrality closer to the Facebook data, however the distribution function was startlingly different.

As can be seen the graphs, the centrality values have very distinct histograms. These differences can be explained in the inadequacies of the random graphs. Because we do not have the ability to recreate how friendships between people are formed, we cannot create a graph similar to the social

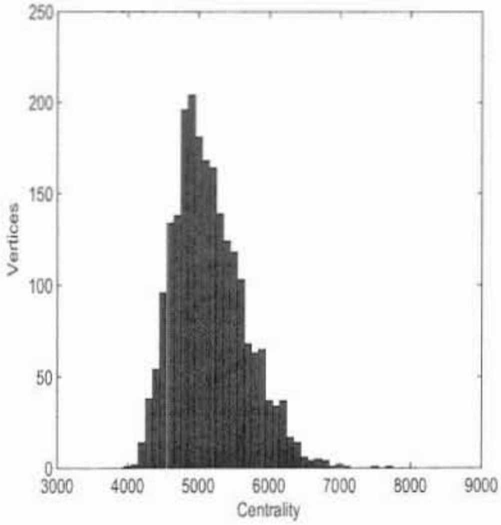


Figure 12: Histogram of Centrality of Facebook Data

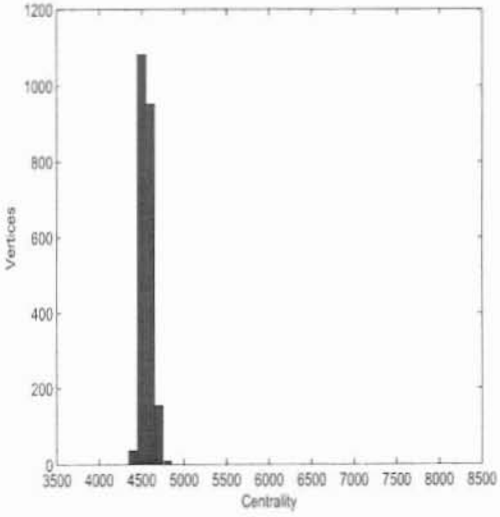


Figure 13: Histogram of Centrality of Random Graph 1

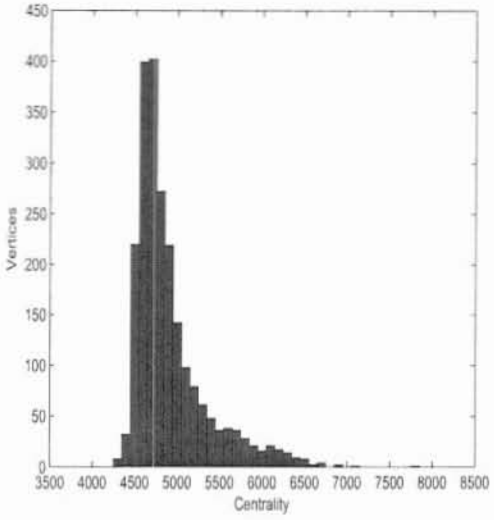


Figure 14: Histogram of Centrality of Random Graph 2

network in regards to centrality. Future models for social networks could possibly be done by trying to create methods where instead of randomly choosing two people to be friends as was done in our first model, or instead of choosing random friends for people as done in the second model, if a more interesting model of friend selection is proposed. Ideas for such a model could include creating artificial cliques in the graph to represent places such as work environments or clubs where people can get together and become friends with each other. Alternatively new friendships can form by being introduced to new people by current friends. Certainly there are other methods as well that could be used to improve the work we started.

7 Club Analysis

So far, we have been primarily interested in showing the data we can glean from an analysis of a friend to friend social network. However, this is not the only kind of social network that we may find on a college campus. We can also analyze the social network created by the interaction of clubs and organizations. Being a member of an organization can both influence the number of friends you have, as well as the connectedness of these friends. Also important, clubs can interact with each other in their own right.

7.1 Sources of Error in our Dataset

Since we are gathering data on club membership from the social networking web site Facebook.com, we must take into account the errors we derive from that source. Of these errors, the foremost among them is inaccurate information entered into the system.

While we chose Facebook.com based primarily on the accuracy of data, this accuracy is limited to ensuring that the person is a member of the school and that the people they claim to be friends with also acknowledge that friendship. It does not extend so far as to give any guarantee regarding the listed activities of its members. Even more importantly, there is no guaranteed formatting of the activities list, so two people who participate in the same events may be listed in completely different ways.

Fortunately, while there is no enforcement regarding the formatting or accuracy of activities, a common convention has been adopted by many Facebook.com users. This convention is to list each activity with a comma to separate them. As such we were able to separate out the vast majority of activities via a simple computer script. The rest we separated by hand, where possible.

Unfortunately, even this list of activities is not as accurate as it can be. As an example, many clubs go by short acronyms. Even more, some people list specific titles or roles in the given organization. It would be near impossible for a script to separate these out, sort them into the appropriate groups, and then recompile a list of members for each activity. In addition, activities may not actually involve some kind of interaction, or may be very vague. A member may, for example, list the activity of tennis, but that does not necessarily mean they are on the tennis team.

In order to try to correct these errors, we first sorted the list of clubs by the number of people in them. We then manually went through the list looking for instances where a club was listed under two different names. After compiling this list of clubs and their possible names, we merged them together so we had one list of hopefully accurate clubs. We also at this point removed any activity that did not involve actual social contact, since our purpose is to study social interaction.

This left us with a list of clubs on which to perform analysis. This list has gone through both automated and manual testing and analysis to ensure it is as accurate as it can be. Unfortunately, no amount of reasonable checking will be capable of ensuring that all activities listed by a person are actually performed by that person, so our data set, while as accurate as possible, still has some errors.

7.2 Types of Analysis

The most interesting form of analysis one can perform on clubs is definitely investigating how being a member of a club relates to your connectivity on campus as well as to the groups other members. For example, we may look at the average centrality of the members of the organization. Alternately, we may just look at the average number of friends each member has. Further, we may look at a groups membership and see how closely connected those members are to each other. Either way, this type of analysis can be useful to prospective members of a group.

However, this is not the only type of analysis one can perform. Just as we looked for centrality and connectivity of people, we can also do the same for groups. In order to do this, we may add an edge between any group that shares at least one member. However, we may go one step further and actually create a weighted graph by using as a weight the number of members the two groups shares. We can easily see that there is no need for a directed graph in this case, since sharing members is clearly a reflexive operation.

7.3 Club Centrality to the Campus

In order to analyze a clubs centrality to the campus, one must first determine that clubs centrality. There are several different ways we could do this, the most obvious being to average the centrality of all the members of that club. However, we felt that this would unfairly bias against clubs who had a large number of central members with a small number of non-central members.

Another way to determine the centrality of a club is to consider all members as one entity for the purposes of determining centrality. The way this works is to create a vertex on our graph such that if there is an edge from any person to any member of the club, there is also an edge to this vertex. We can then check the centrality of this person, and use it as our club representation. We felt that this method would most accurately represent a club's centrality to the campus body, and as such it is the method we went with.

There are, of course, problems with choosing this method. The largest of these problems is that it may appear to impart an unfair advantage to larger clubs. In this method, we only consider the shortest path to any external vertex from any member, so a club can only benefit from more members, not be harmed. This means that membership has a monotic implication towards club centrality. However, this is often in fact the way clubs work in reality. A club having a member with few friends does not frequently impact the influence of the club as a whole, since his lack of connectivity does not impact the connectivity of other members of the club.

7.4 Clubs and other Clubs

While club centrality to the average student on campus is an important metric, it can also be useful to think of the implications of how a club relates to other clubs. Once again, we can view clubs as a graph. In this case, however, instead of having an edge represent a friendship between two nodes, we will define an edge as representing a shared member between two clubs. This will result in giving us a simple graph of relations between clubs.

7.4.1 Co-sponsored Events

One of the things we can look into is whether clubs that share members co-sponsor more events with each other. A co-sponsored event is one in which more than one club takes responsibility for running the event. These events can be larger in scope than many clubs could run on their own, but they also involve more difficulties with communication. Our belief is that if two clubs share many members, they would have an easier time communicating and thus be more likely to co-sponsor an event. Of course, as the union of the clubs increase the size the benefit of co-sponsorship diminishes due to overlap in members.

7.4.2 Club Centrality to other Clubs

We looked earlier at club centrality to the campus as a whole. However, this is not the only centrality we can consider with regard to clubs. Also useful to know is how central a club is relative to other clubs. If we consider centrality in general as a representation of both how well known and how easy it is for a person to get a message across the network, then we can see that a more central club can have an easier time of both running events as well as informing other clubs (and thus their members) about them.

It is our belief that there is a relationship between being central on campus and central to other clubs. This is to say that if a club is not central to the campus as a whole, it will also be very difficult for it to be central to the clubs that exist on that campus. Logically, how could a club be central to an organization when it is not central to that organization's members?

In order to determine centrality, we must first decide how to determine whether two clubs are connected to each other or not. Originally, we planned on saying that if a club had a shared member with another club, those two clubs were connected. Since sharing members is a symmetric operation, the graph would not need to be directed. However, this seemed to give both too much weight to one person being in two clubs and too little weight to two clubs having many members in common.

We also considered using a weighted edge to represent club connectivity. In this case, we would weight the edge based on how many members the two clubs share. This solves the problem of ignoring the number of shared members, however we felt it was still an inaccurate representation. For one thing, it seems unfairly biased towards large clubs, which would most likely end up being central just due to their size. However, the more glaring inaccuracy was that this method completely ignored the possibility that a member of one club may be close friends with a member of another.

In order to solve this problem, we came up with the following means of determining club connectivity. First, we decided that all clubs were connected in some way, so there should be an edge between every possible pairing of clubs. However, we also decided that the strength of that connection should be based on the strength of the club's member's connectivity to the members of the other club. Since this may mean that club A is better connected to club B than club B is connected to A, our final result needs to be a directed graph. The final calculation we used to determine the strength of this calculation was the average of the minimum distance from each member of club

A to any member in club B.

As an example, let us consider the following two clubs. The first club is called *Really Fun Club* and the second club is called *Also Fun Club*. In this example, Really Fun Club has three members, Sam, Bob, and Mary. Also Fun Club has only two members, Mary and Joe, for a total of four different people considered. Of these four people, the following friendships exist. Sam is friends with Bob and Joe and Mary has no friends. As we found out later in our research into gender, this is not realistic, however it works for this example.

To start this example, we must first determine the minimum distance between the people that make up these clubs. These can be represented by an adjacency matrix, shown in Table 2. Once we have the minimum distance between any two people, we can then construct a new matrix. In this matrix, we will list the members of one club for the rows, and the members of the other as a column. The cells of the matrix will be the minimum distance between the two members, resulting in Table 3.

From here, the calculation is simple. We take the minimum value of each row, add them up, and take the average. This represents the distance from Really Fun Club to Also Fun Club. We then do the same for the columns, representing the distance from Also Fun Club to Really Fun Club. The results of this can be seen in Figure 5, with Really Fun Club having a distance of 1 from Also Fun Club and Also Fun Club having a distance of $\frac{1}{2}$

After we've found the minimal path between all possible pairings of clubs, we simply used the same algorithm for computing centrality for people on clubs. The results were, in this case, surprising. For our graph, there was no real correspondence between a club's centrality to the campus as a whole and a club's centrality to other clubs, as can be seen in the scatter plot figure. This is counter-intuitive, given our definition for weighting of a club's edges, but may be explained by the fact that we consider the distance from a club to another club to be a function of the minimal path of each person in the first to the second, as opposed to the average path. This can cause a club who has a single friend in many clubs to raise the centrality of any club that person is in drastically, resulting in the results we saw.

Table 4: The adjacency matrix for our example

	Sam	Bob	Mary	Joe
Sam	0	1	∞	1
Bob	1	0	∞	2
Mary	∞	∞	0	∞
Joe	1	2	∞	0

Table 5: The adjacency matrix of club members

	Mary	Joe
Sam	∞	1
Bob	∞	2
Mary	0	∞

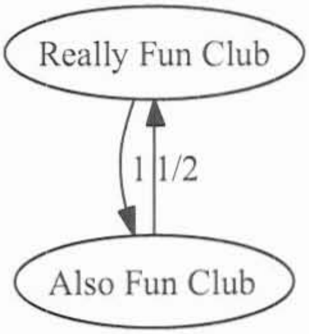


Figure 15: A Weighted Graph Representing Club Distances

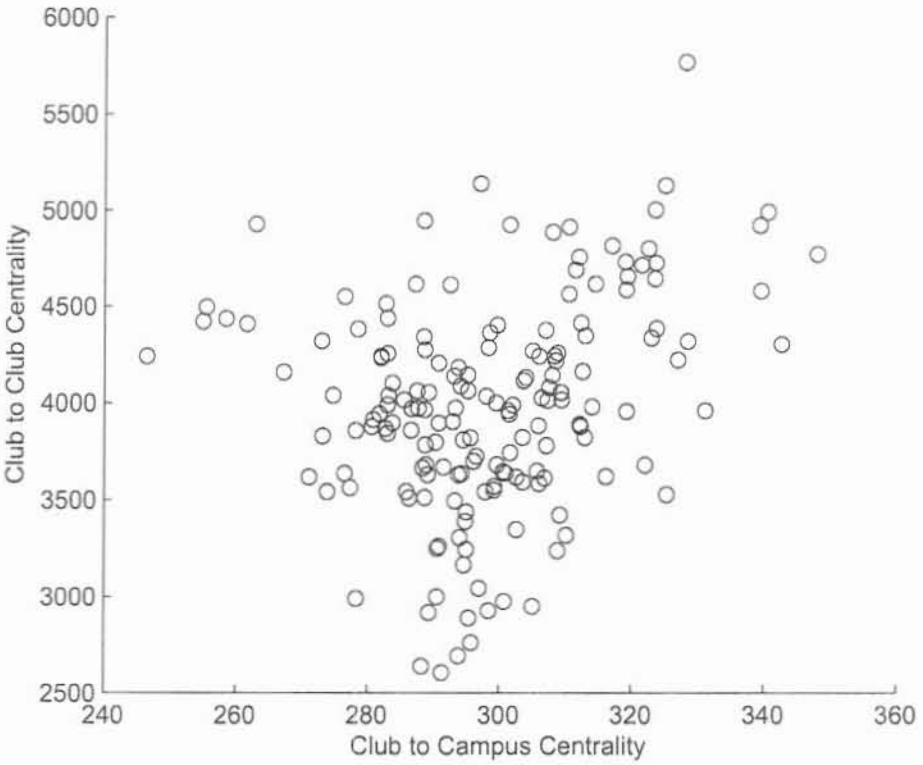


Figure 16: A scatter plot of club centrality calculation methods

8 WPI Student Government Association Analysis

One of the uses for the graph of WPI students is to analyze the effectiveness of the WPI student government. The WPI Student Government Association (SGA) is a group of approximately 40 students who determine policies and programs for the student body. The idea of the student government is that the students involved should be an accurate representation of the students at WPI. The students on SGA should be a model of the entire campus, so different opinions and ideas from around campus are heard in the government. The way student government elections work is that for every 80 students living on-campus there is one seat on the senate for a person who lives on campus. For every 80 students living off-campus there is a seat on the senate for a person living off campus. People living on campus can only vote in the elections for people living on-campus, while people living off-campus can only vote for off-campus seats. So the on and off campus elections are independent of each other. Within each election, the students are allowed to vote for as many people running as there are seats in the election. So if there are 30 people running for off campus seats and there are 20 off campus seats available, a student living off campus has 20 votes to split between those 30 students. Any one person can only vote for a candidate once, so actions such as putting all 20 votes into one candidate are not allowed.

Using our graph we can model this behavior of voting and see if the winners accurately represent the campus body. For the purposes of our model, we consider a student to be represented on the Student Government Association if someone on SGA knows them. To create the model, we looked at several behaviors inherent in the voting patterns from SGA elections in the past. First, people tend to only vote for candidates they know. This is because candidates tend to tell all their friends to go vote when elections are taking place, and thus when the students go to vote, they vote for the candidates that they have talked to. While students do occasionally vote for people based on the brief description on the ballet, we could not accurately model the many factors that go into such a decision. Also, students do not tend to use all their votes if they do not want to; if they have several potential votes left but do not want to vote for any of the other candidates, they will not. Finally, in the senate elections only about 30 percent of the student body votes, while the rest just ignore the election. We believe that the people who do vote are the people who know someone running and have told them to vote. Using these rules, we can build a model of the elections process.

Our election model works as follows:

1. Randomly choose n/40 students to be candidates, where n is the number of students in our model. These students will be running for n/80 seats.

2. For each person in our dataset, run a voting algorithm. The algorithm works by first determining randomly if the student even votes in the first place. Then it checks if the student knows any of the candidates in the election. If they do then they will cast a vote for someone they know. If they end up knowing more of the candidates than they can vote for, randomly choose which of the candidates they know to vote for.

3. Tally up all of the votes and determine the winners.

Unfortunately, our model is flawed in that our data does not tell us who is living on-campus and who is living off-campus, so to overcome this we allowed anyone to vote for anyone, and made no distinction between on and off campus students. This flaw should not invalidate the study, since the method of voting for candidates that people tend to know is the same, the only difference is that there are no artificial boundaries set up by the election process.

Before running this model, we hypothesized that the SGA would not be an accurate representation of the entire campus. We believed this to be the case because using this model, the candidates who know many people tend to get the highest number of votes, and thus represent all of their acquaintances on the Student Government Association. However, the students who are mostly friends with people in smaller cliques will not know these high degree candidates, and any candidates they vote for will be beaten by the people with higher degree. So the Student Government Association will be filled with students who belong to large cliques, and the smaller cliques could not muster enough votes to get any representation.

The model was run 10,000 times on the data, and the results were startling. Only around 63 percent of the campus will be represented by this method of voting. The standard deviation of this data set was only 0.03 percent, which shows that regardless of who decides to run, only 63 percent of campus can be represented with this number of people. This data shows us that there are many people who are simply left out of the student government, with little form of representation on campus. Also, if the number of seats on the Student Government Association Senate is doubled so that for every 40 students on campus there is a seat on the senate, then our model showed that 79 percent of the campus would be represented. So in terms of representing the campus, it might be better for the WPI SGA to consider increasing the amount of senators in the group. This doesn't take into consideration any problems with the logistics of running a senate of double the size, merely how to better represent the campus.

At other colleges student government have different methods of elections. One way that student governments are formed are to have each club on campus send a representative to the meetings. In that manner, each club elects its own "senator" to represent it on the student government. Using our data we were able to take our club lists and model this election system.

Our first model took our 171 clubs and found the person with the most friends in the club and had them be the representative. The idea here is that the club will have their most well connected person be their club representative since they would most likely know a lot about campus. With these 171 people, 92.75 percent of campus would know someone on the SGA. If we were to apply the WPI method of voting with 171 people on the senate, we would get a 93.57 percent coverage of the campus. Therefore the method of having club representatives is not very good for the campus.

Our second club based model was designed in a way such that instead of clubs putting their most popular person on the senate, they put a random club member on the senate. Using that model, the 171 person senate would only represent 65.52 percent of campus. That percentage is hardly an improvement on the coverage achieved by 40 people using the current voting method.

This leads to the interesting question of what would be the minimum size of the senate such that 100 percent of the students at WPI were directly connected to someone on the student government. For this we can choose from anyone in our graph to be hand picked to be on the senate. This directly correlates to the least dominating set problem, which is to find the minimum set of vertices in a graph such that all points in the graph are either in the set, or a distance of one away from at least one of the points in the set. Unfortunately, this problem was found to be NP-Complete [10].

An NP problem is one that can be computed in polynomial time using nondeterminism. Additionally, an NP-Complete problem is an NP problem that has the additional property that all problems in NP are reducible to it. Unfortunately, no algorithm has been found that can reduce an NP-Complete problem to one that can be solved in polynomial time deterministically, and as such NP-Complete problems can only be solved in exponential time currently. Many computer scientists believe that no such algorithm exists, but this has not yet been proven. When the graph contains 2,234 vertices, that quickly renders the problem unreasonably difficult. We attempted to find the dominating set using a program coded in Matlab on a Intel Core 2 Duo processor and found no result after four straight days of processing.

There are methods to approximate the dominating set problem. One such approximation is a greedy algorithm. The algorithm works by finding the vertex with the highest degree and putting it in the dominating set. The algorithm then finds the vertex has the highest degree of connections to vertices that are not already connected to someone in the dominating set, and puts that vertex into the dominating set. This step is repeated until every vertex is connected to at least one vertex in the dominating set. This algorithm works in polynomial time and does find a dominating set. However it does not necessarily find the smallest dominating set. On our graph, we found that a dominating set of 76 vertices exists in the graph. To put it in social terms, there are 76 people on campus that when all put on the senate, everyone who is in our test group will know someone on the senate. There may be a smaller group of people than the 76 we choose, but it can be done with at least 76 people.

With this data, we can conclude that the current voting method for the WPI Student Government Association is not optimal. Increasing the size of the WPI senate, as well as creating ways to get senators who are connected to the peripherial of the campus may be a better way of choosing potential senators. However, the WPI SGA should not consider having clubs choose representatives to be on the senate, since that at best matches the current voting system and at worst has less coverage with far more people. Unfortunately, we have not found any conclusive evidence about any better voting methods for the WPI Student Government Association.

9 Gender Analysis

The data we took from Facebook.com also contained information about the genders of the users. A person's gender was determined by what they listed their gender as on the site. The listing of gender is optional, and as such we do not have the genders of everyone involved in our study. We were able to find out the genders of 2109 people out of the 2234 total in the graph. We felt that having data on the gender of 95 percent of the people involved was acceptable to have minimal error.

Of the 2109 people that listed their gender, 642 of them were female and 1467 of them were male. So of the people who listed their gender, 30 percent were female and 70 percent were male. The official WPI record states that 24 percent of undergraduate students on campus are female and 76 percent are male. We can derive from this that females at WPI are more likely to use Facebook than males are. However without a hard figure as to the number of total students who could possibly use Facebook (this includes graduate students and former WPI students) we cannot figure out precisely how much more likely females are to use Facebook. Although this data has a possible error source from the fact that the data as to the ratio of men/women on campus of WPI is based on current students, so alumni are not taken into account in this ratio.

In our graph, the average female has 102.7 friends, while the average male has 64.0 friends. Females are also more centralized in the graph of WPI than males are, having an average centrally that is of lower value than the average male by calculating the method proposed by Bonacich and the method proposed by Sabidussi. Other interesting data about gender can be found through the graph as well. Consider ranking vertices by their degree. The ninetieth percentile of this group is 62 percent female and 38 percent male. This completely goes against the male to female ratio of people in this study, leading to the conclusion that the people with the highest amount of friends are female. Similarly, of the ninetieth percentile of vertices regarding centrality, 57 percent are female and 43 percent are male.

10 Conclusions

In conclusion, our analysis has turned up many interesting facts about the WPI campus. We have discovered that the average distance between any two members is low, with few outliers. As such, members on our campus are connected relatively tightly to one another. We have also discovered that Social Networks do not form along discernible mathematical lines. Simulating the creation of one can be difficult, and though it is possible to match part of the data patterns this does not necessarily guarantee that we will match the overall effect. With regard to clubs, we have shown that a club can be central with respect to the campus and yet still not central with respect to other clubs.

Additionally, we have discussed representation in a student government, and how guaranteeing optimal representation is a problem that is computationally expensive to solve. We have shown that the given voting model is most likely inadequate, representing approximately $\frac{2}{3}rd$ of the campus. We have also provided other possible ways of determining senators, though we have not discovered one guaranteed to provide fair representation. Finally, we have also discussed the effects one gender can have on their social experience on the WPI campus.

References

[1] http://www.alexa.com.

[2] http://www.facebook.com.

[3] http://www.livejournal.com.

[4] http://www.myspace.com.

[5] BONACICH, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology 2* (1972), 113–120.

[6] BORGATTI, S., AND EVERETT, M. A graph-theoretic perspective on centrality. *Social Networks 28* (2006), 466–484.

[7] BUCKLEY, F., AND LEWINTER, M. *A Friendly Introduction to Graph Theory.* Prentice Hall, 2002.

[8] CHARTRAND, G. *Introductory Graph Theory.* Dover Publications, 1984.

[9] CORMEN, T., LEISERSON, C., AND RIVEST, R. *Introduction To Algorithms.* The MIT Press, 1997.

[10] GAREY, M., AND JOHNSON, D. Computers and intractability: A guide to the theory of np-completeness, 1979.

[11] ROSS, S. *A First Course in Probability.* Prentice Hall, 2005.

[12] SABIDUSSI, G. The centrality index of a graph. *Psychometrika 31* (1966), 581–603.

[13] WASSERMAN, S., AND FAUST, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.