

Project Number:46-KXF-IM02

AIMBASE:
A SOCIAL STUDY OF INSTANT MESSAGING

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

by



E.O. Stinson

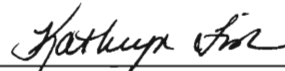


Jeffrey A. Rosenberger

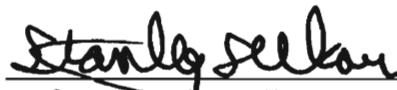
Date: May 1, 2003

Approved:

1. social networks
2. instant messaging
3. graph theory



Professor Kathi Fisler, Major Advisor



Professor Stanley Selkow, Co-Advisor

Table of Contents

1 –	Introduction.....	1–7
2 –	Instant Messaging.....	2–8
2.1 –	Using AIM for the Study of Instant Messaging	2–9
3 –	Background.....	3–10
3.1 –	Previous Studies	3–10
3.1.1 –	The “Small World” Problem	3–10
3.1.2 –	The Erdős Number Project.....	3–11
3.1.3 –	The Oracle of Bacon	3–12
3.2 –	Literature Review	3–13
3.2.1 –	Social Networks and Six Degrees.....	3–13
3.2.2 –	Network Structure vs. the Propagation of Information.....	3–14
3.2.3 –	Methods of Communication.....	3–16
3.2.4 –	Groups and Context	3–17
4 –	The AIMBase Experiment: Goals and Methodology.....	4–19
4.1 –	Questions.....	4–19
4.2 –	Graph Generation.....	4–20
4.3 –	Survey Execution.....	4–20
4.4 –	Finding Cliques.....	4–20
5 –	Initial Experiment Results: AIMBase as a “Small World”.....	5–22
6 –	Survey Results and Analysis.....	6–24
6.1 –	Geographical Clustering.....	6–24
6.2 –	AIM and Face-to-Face Interaction.....	6–30

6.3 –	The Effect of Instant Messaging on Phone and Email Usage	6–32
6.4 –	Information Propagation	6–34
6.5 –	Instant Messaging Productivity	6–37
6.6 –	Demographics of AIM Users	6–41
6.6.1 –	Age	6–41
6.6.2 –	Gender	6–43
6.6.3 –	Geography	6–44
6.7 –	Cliques and Demographics.....	6–45
7 –	Technology.....	7–50
7.1 –	Web Server	7–50
8 –	Conclusions.....	8–51
8.1 –	Sample Size Disclaimer	8–52
8.2 –	Recommendations for Further Research.....	8–52
9 –	Additional Information	9–54
10 –	References.....	10–55

Table of Figures

Figure 3-1: Distribution of Bacon Numbers.....	3—12
Figure 5-1: Distribution of AIMBase Maximum Distances.....	5—23
Figure 6-1: Distribution of Link Distances (Home-to-Home).....	6—26
Figure 6-2: Distribution of Link Distances (College-to-College)	6—27
Figure 6-3: Distribution of Link Distances (Minimum).....	6—29
Figure 6-4: Online vs. Face-to-Face Relationships.....	6—30
Figure 6-5: The Lattice Model.....	6—31
Figure 6-6: Changes in Email Usage Due to AIM.....	6—33
Figure 6-7: Changes in Phone Usage Due to AIM	6—34
Figure 6-8: Disease Front Propagation.....	6—34
Figure 6-9: Log Distribution of Links in Database.....	6—35
Figure 6-10: Links Added Per Day (Hub Discovery).....	6—36
Figure 6-11: Survey Respondents Using AIM for Business Purposes	6—38
Figure 6-12: AIMBase Participant Occupations.....	6—39
Figure 6-13: Survey Respondents Using AIM for Academic Purposes	6—40
Figure 6-14: Survey Respondents Using AIM for Increased Productivity	6—41
Figure 6-15: Harris Poll of Ages of Adults Online.....	6—42
Figure 6-16: Ages of AIMBase Survey Respondents.....	6—42
Figure 6-17: Gender Divisions of AIMBase Participants	6—44
Figure 6-18: Geographical Distribution of AIMBase Participants' Home Addresses	6—44
Figure 6-19: Geographical Distribution of AIMBase Participants' College Addresses	6—45

Figure 6-20: WPI Clique (Dense Subgraph #1)..... 6—47

Figure 6-21: 7-Clique (Dense Subgraph #2) 6—48

Abstract

This project examines the social network of instant messaging (IM) systems. We studied the connections within the America Online Instant Messenger community both socially and graphically. Major areas studied include the geographical clustering of IM communities, demographics, the spread of information through the IM network, and the effects of IM on productivity and on the use of other communication systems. The direct results of this study are applicable to marketing, workplace productivity, and social psychology.

1 – Introduction

This study is an attempt to discover new information about the ways in which computer mediated communication occurs via synchronous (“instant”) messaging. We focused on investigating the ways in which communities of users interact, with regard to acquaintance relationships, geography, age, and other demographic factors of interest. Our study focuses almost entirely on the AOL Instant Messaging (AIM) network, which includes more than 180 million users (*AOL Announces Expansion*), and is widely accepted in many environments. Our intent is that knowledge gained in this analysis will be used to provide insight into sociological group formation in online environments, especially through synchronous communication mediums. Like much other social communication research, the direct results of this study are applicable to marketing (in terms of data mining and community profiling), workplace productivity (as many businesses are beginning to allow or promote internal instant-messaging networks), and psychological studies (by examining inter- and intra-group interactions).

2 – Instant Messaging

Various means of interpersonal communication have been the driving force behind adoption of the Internet. In the early days of the Internet/ARPANet, email was the “killer app” of the network. Email was important in proving the worth of the Internet to non-technical users because it was an application everyone could understand and use. Other uses for the network at the time were only useful to those with an interest in computing. Email gave people an ability to communicate without the costs and synchronous nature of the telephone. Due to the asynchronous nature of email, one user can send a message that will not be received or read until later, much like regular mail. Unfortunately, this delay could make the turnaround speed of email conversations rather slow compared to telephone conversations.

Now, synchronous communication is making a comeback as computers become more easily accessible during everyday life. In recent years, so-called “Instant Messaging” systems (often referred to as IM systems) have become very popular. These systems allow for synchronous communication, much like the telephone. As soon as one user sends a message, the recipient is aware of it, and has an opportunity to respond. There were various synchronous systems available in the past, such as IRC and “talk mode”, but these did not become very popular with non-technical users because of the level of expertise required to use them.

In 1996, what is widely considered the first popular IM client, called ICQ (for “I Seek You”) was developed in Israel. It became hugely popular until the company was eventually bought by AOL Inc. in order to reduce competition with the at-the-time newly introduced AOL Instant Messenger. (Saunders) The service was kept running, however,

and now ICQ has 140 million registered users. (*AOL Fast Facts*) The AIM and ICQ networks are also being integrated in a recent move by AOL to confirm their status as the instant-messaging leader.

2.1 – Using AIM for the Study of Instant Messaging

With its 140 million registered users, the most popular instant messaging system is now AOL Instant Messenger, also called AIM. Our study will focus on AIM, but should be applicable to other instant messaging systems provided by other companies. AIM has several features that make it more useable by non-technical users than previous systems, such as IRC. AIM maintains a personal “buddy list”, which allows users to enter a list of people they communicate with frequently. The AIM client software displays the buddy list, and indicates which of the people on it are currently connected to the AIM network. AIM also has a profile feature. A profile is a brief piece of text created by the user and available for easy viewing by other users. Profiles are generally used to contain song lyrics, web links, quotes, and other brief material, and can contain some web-like content including links and formatted text.

Some studies of online communities and social networks have already been done, mainly using email systems. The instant messaging communities that we are studying have been mostly ignored due to the problems of data collection inherent in systems where all centralized communication is proprietary. Previous studies also generally focus on professional relationships; our study looks at both personal and professional relationships as much as possible. We attempted to collect data that would let us find correlations demographics, AIM usage, and the spread of information over instant messaging networks.

3 – Background

3.1 – Previous Studies

3.1.1 – The “Small World” Problem

The phrase of “six degrees of separation” comes from the research of Stanley Milgram, in his study “The Small World Problem”, published in 1967 in the journal *Psychology Today*. (Milgram, Greengard) A target person (a stockbroker) was chosen in Cambridge, MA, and then 300 folders were distributed to three different groups of people. The first hundred were in the Boston area, the second hundred were blue-chip stock investors in Omaha, and the third hundred (also from Omaha) were chosen randomly from a mailing list. There were rules regarding the handling of the folder (for example, each person who held the folder was required to write their name and location in the folder). The most important rule required that the folder could only be passed on to someone that the current keeper of the folder knew on a first-name basis. The general idea was that the folder would follow a path of close acquaintances until it reached the target person. Since each person who held the folder was recorded, the project coordinators could track the number of steps from start to finish. It was discovered that the median number of steps was 5, so that, counting the end person, six people were involved, hence the phrase “six degrees of separation”.

Milgram’s research has come under question, however. Some groups, including a team at the University of Columbia, point out that his research was very limited and his claim that almost everyone is connected within six steps should not be used as a general

rule of thumb. Many people who study social networks take his conclusions for granted, despite the fact that only a third of his starting subjects were completely random and, of those, only 16 of the folders reached their destination! In an effort to verify Milgram's conclusions, the Columbia University team is attempting to recreate his experiment using email. (Columbia University) Similarly, the AIMBase project is an attempt to discover connectedness across the AOL Instant Messenger network.

3.1.2 – The Erdős Number Project

The Erdős Number Project, maintained by Jerry Grossman and others, is a project to create a linked graph of related mathematicians. Paul Erdős was an incredibly prolific Hungarian mathematician who participated in many collaborative publications with other mathematicians. By assigning him an Erdős number of 0, and his immediate collaborators a value of 1, and their immediate collaborators a value of 2, etc., it is possible to be a graph connecting almost all mathematicians. For example, Andras Sarkozy collaborated with Erdős on at least 62 papers, so has an Erdős number of 1. Sebastien Ferenczi worked with Sarkozy, so has an Erdős number of 2. Unlike Milgram's "Small World" project, which focused on the sociological aspects of connectedness, the Erdős Number Project is run by and for mathematicians, and focuses on the way community networks form from a graph-theory perspective rather than on the interpersonal relationships between mathematicians. Admittedly, there are some useful measures of a graph that relate to how people interact, which will be covered later. (Grossman)

3.1.3 – The Oracle of Bacon

The Oracle of Bacon, at the University of Virginia, appeared in response to the Kevin Bacon Game. The Kevin Bacon Game is a popular parlor game based on finding the relationships between the popular actor Kevin Bacon and other actors. Instead of using collaborations based on published mathematical papers, the game uses film collaborations. If an actor has been in a film with Kevin Bacon, then they have a Bacon number of 1, and so on. The Oracle of Bacon is a web site that utilizes data collected by the Internet Movie Database (IMDb.com) in order to programmatically find relationships between actors. The website also provides the ability to search for relationships between any two actors, or to other famous actors (including Arnold Schwarzenegger and Elvis). Asking the Oracle of Bacon how Daniel Radcliffe, the star of the Harry Potter movies, is related to Bacon produces the following response:

The Oracle says: Daniel Radcliffe has a Bacon number of 2.

Daniel Radcliffe was in Harry Potter and the Chamber of Secrets (2002) with Miriam Margolyes

Miriam Margolyes was in Balto (1995) with Kevin Bacon

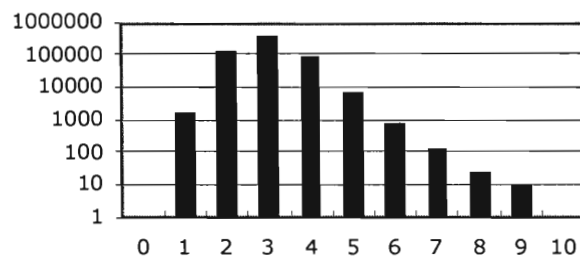


Figure 3-1: Distribution of Bacon Numbers

It is worthwhile to note that the Oracle only produces the first minimal path between two actors that it comes across. Figure 3-1 shows a logarithmic chart of the distribution of Bacon numbers. On the horizontal axis is the number of steps to Kevin Bacon, and on the vertical axis is the number of actors. Note how there are very few actors with a Bacon number greater than six. As of March 26, 2003 (when this graph was generated), there were only 152 actors removed by more than six steps. Compare this to the 559169 actors within six degrees of Kevin Bacon.

The Bacon graph is also interesting in that restricting “interaction” to mean just the collaboration of two actors in a movie still produces a graph of degrees of separation that seems to reinforce Milgram’s theory that everyone is within six degrees of everyone else. Admittedly, choosing as prolific an actor as Kevin Bacon makes for slightly closer numbers. There is a link from the Oracle of Bacon site showing the person who is the worst “center” for the universe of actors, and it takes around thirteen steps to find most of the actors connected to this person. We can explain this seeming problem by noting again that defining interaction as collaboration in movies is a fairly restricted view. (Reynolds)

3.2 – Literature Review

3.2.1 – Social Networks and Six Degrees

Social network theory is an attempt to explain what appear to be diverse communities and networks using a common set of theoretical tools. It is now commonly accepted that almost any system can be reduced to a type of social network for easier study and analysis. One of the most social of the social networks is the interaction and

communication between people; other social networks including everything from the way in which crickets synchronize their chirps to the fluctuations in highway traffic. By discovering some of the attributes of the AIM network, we can make conclusions about the ways users interact. In order to describe our findings, it is useful to understand some of the theory behind the study of social networks.

3.2.2 – Network Structure vs. the Propagation of Information

Two broad approaches can be taken when exploring social networks. The first is to examine the structure of the network in comparison to the underlying societal groupings. This branch of study attempts to infer information about the people who are part of the network based on the shape of the network. Consider the simplified example of a network that seems to contain two fairly separated groups. Assume that the data was collected from all college students in a city. Further study indicates that each group loosely represents students at one of the city's two different colleges. The connections between the two are people at one college with friends at the other. Given a little bit of information about the collection of data, it becomes possible to make useful conclusions about relationships.

Discrete groups within a network are generally referred to as *cliques*, and mostly-discrete groups (where some connections between the two occur, as in the example) are called *partial cliques*. Later in this report we show how well AIM users form partial cliques based on geography, interest groups, and other factors.

The second way of looking at social networks is to pay attention to the travel of information, rather than the links. This approach is, as Watts describes it, more

“mechanistic”. (Watts) This is the study of how information travels through a social network. (Think of it as the study of social “grapevines”.) In this case, we focus not only on what groups a person is part of, but also on his or her position within those groups, and the information within those groups.

People with many links to others, called *hubs*, are more likely to know something than those with few social connections to other people. For example, a travel agent is part of the group of people trying to sell plane tickets. They know most of the airline salespeople that are also trying to sell plane tickets, whereas each airline salesperson only knows about their own tickets. When someone is looking for the cheapest ticket available, the travel agent is the obvious person to ask. It became obvious early on that certain users of AIM act as hubs, with connections to many other users. Because of this, popular ideas (*memes*) can potentially travel very quickly across the instant messaging networks. If a popular (hub) user talks about something to most of their acquaintances, the information gets a sudden boost in popularity. Since this happens every time that a new hub person learns a meme, it can be argued that they support the small-worldness of the AIM network.

The example of the travel agent illustrates another interesting feature of social networks. We say that two people in a network are *strongly tied* if they are part of the same social group and/or interact frequently with each other. In the example above, the travel agent and the airline salespeople are strongly tied. On the other hand, the travel agent and his customer are *weakly tied*, meaning that they only interact infrequently. Somewhat surprisingly, it turns out that weak ties are more important for effective social coordination than strong ties. With a closer look, this makes more sense; it is not the

people we spend lots of time with that give us new information and opportunities, but the casual acquaintances we have in the right places. The “strength of weak ties”, as it was called by Mark Granovetter, the first person to study the phenomena in depth, is an important concept in understanding how information moves around in a social network. (Watts, pp. 48-49)

3.2.3 – Methods of Communication

In Milgram’s study, each of the starting subjects was given one folder. They were then asked to pass that folder on to one person who they thought was “closer” to the destination person. This is an example of a directed search, in which people must choose the person they think is closest to the target. The other type of search is called “broadcast”, and occurs when instead of passing along the folder to just one person, every friend is contacted. Directed searches are studied very frequently by people interested in social networks, partially because they are generally much easier to study. Imagine if Milgram had asked the people in his experiment to make a copy of the folder and pass it on to *everyone* they knew, instead of just the person they considered most likely to be on the path to the target. By the end of the experiment, 300 folders would have reached every person on the planet, hardly a feasible way to conduct an experiment.

Interestingly enough, our experiment, by its nature, is based around the broadcast of information, albeit in a more apathetic way than distributing folders. In the digital world of AIM, all it takes for someone to tell everyone that knows them that they are part of the AIMBase project is the simple act of adding some text to their profile. There is no need to manually make dozens of copies; the medium does it for us. In this way, each

person who joins the AIMBase project is automatically broadcasting that fact to anyone who reads his or her profile.

In fact, much communication in the digital world occurs in this way. Web pages do not just pass information on to one person; they pass it on to every visitor. Message boards and chat rooms also provide a medium for broadcast communication. This explains in part why ideas seem to rush back and forth over the Internet so quickly.

3.2.4 – Groups and Context

How do groups affect how people interact? This is one of the most important questions of social networking. Relationships are not formed between people randomly, or independent of their current social environment (their “context”). Instead, people are part of groups, and these groups determine how people interact with people they already know, and how they meet new people. Most people are part of several groups, and those groups may interact very little; the people you know from a project at work may be almost completely separate from the people you’ve known since high school. For everyone in the two groups, finding that you are the link to the other group may not be easy, though eventually your presence will probably lead to at least a few more links between the two groups.

It turns out that groups are, in fact, the glue that holds the small world together. It has been shown, both theoretically and empirically, that when people are asked to participate in a study like Milgram’s, they generally use a standard mode to determine whom they will pass the folder to, so to speak. (Watts pp. 134) From their acquaintances, the groups they are involved in, they choose the person who is most likely to be in a

group with the target person, or a person close to the target person. This pairing of groups in order to determine connections is a deep part of our social structure, and explains the way in which a piece of information can travel across the globe from one person to a seemingly completely unconnected person in so few steps. (Watts pp. 71)

4 – The AIMBase Experiment: Goals and Methodology

4.1 – Questions

As part of our initial project proposal, we came up with the following set of questions. We managed to come up with satisfactory answers to all except the first question. We decided that finding interest groups would require a survey that would be more invasive than most people would accept.

1. How much cross-community interaction is there? Do AIM users only interact with people within a set community, or are they more likely to be members of many communities?
2. How do memes spread through online communities, specifically the AIM network? What does this imply about levels and types of information exchange?
3. How much is AIM used in work or school environments for exchange of “productive” information? In other words, is AIM a useful productivity tool, or just a social facilitator?
4. Demographics:
 - a. How do age, gender, etc. affect AIM usage?
 - b. How do colleges or other geographically discrete communities affect community formation?
 - c. How do interest groups affect community formation?
5. What correlation is there between AIM interaction and interaction in real life?
6. To what extent is AIM used as a replacement for other methods of communication?

4.2 – Graph Generation

When an AIM user clicks on the AIMBase link in a user's profile, our server records the two AIM names: the clicker of the link, and the clickee who had the link in their profile. These pairings of users are recorded in a database, along with the time when the action occurred. We use this data to generate the graph of AIMBase users and to find cliques (mentioned in detail later) and other information.

4.3 – Survey Execution

We conducted the project survey by using an AIM 'bot written specifically for the project. The 'bot selects a subset of users from the database and adds them to its internal buddy list. When one of these users is online and not idle, it sends them a message requesting that they take the survey. The 'bot will only message people once per day. When a user goes to the survey page, they can either take the survey or request that they never be contacted again.

4.4 – Finding Cliques

Social network theory commonly describes networks from the point of view of graph theory. The people or entities within the graph are called *nodes*, and the connections between them are called *edges*. A dense subgraph is a subset of a graph where the nodes have many connections to other nodes within that same subset. We wrote a software tool to pull out dense subgraphs so that we could compare the density of connections to the demographic relationships of people. For example, we expected people in close

geographic proximity to be more densely interconnected than others. In Section 6.7 – Cliques and Demographics, we discuss how the subgraphs discovered in this process compare to our survey results.

Cliques are subgraphs that are completely connected; meaning everyone in the clique is connected to everyone else in the clique. Rather than finding complete cliques, we decided it was sufficient to find densely connected subgraphs that are almost cliques. Finding cliques is an NP complete problem, so rather than exhaustively searching for them, we chose the following heuristic.

The software written to pull densely connected subgraphs out of the AIMBase graph has some shortcomings, but it still provided some interesting data. The algorithm works like so: Given a threshold value of n , we remove all of the people from the graph who are directly connected to less than n other people. Since there are now fewer people in the graph, some people who had n links before now have less than n , so we repeat the algorithm until no more people are removed. At this point, we have a set of disconnected or loosely connected subgraphs. In Section 6.7 – Cliques and Demographics, we discuss how the subgraphs we discovered relate to the real-life relationships between AIMBase participants.

5 – Initial Experiment Results: AIMBase as a “Small World”

In order to encourage AIM users to participate in the AIMBase project, our website provides an interface for discovering connections between users much as the Oracle does for actors. We used a very similar algorithm to discover distances between users. By computing the average distances to each person in the AIMBase, we can also discover the best and worst centers of AIM users in our database. As of the writing of this report, the average number of steps from the best center to any other participants in the project is 3.3. Even more exciting is that the maximum number of steps from the best center (and 7 other people) to anyone else in the graph is only 8. We expect that this number is two more than the commonly accepted 6 degrees because our data does not represent more than a small fraction of the AIM network, and the boundaries of the graph (formed by people who have recently joined the project) are not very complete.

When we locate the “worst” centers, we discover that the average number of degrees between the worst center and anyone else is 10.8, and the maximum number of steps they have to take to reach another person is 15. According to the Oracle of Bacon website, the worst center in the universe of actors is at most 16 steps from anyone else. According to our data, the AIM network is more closely connected than the universe of actors.

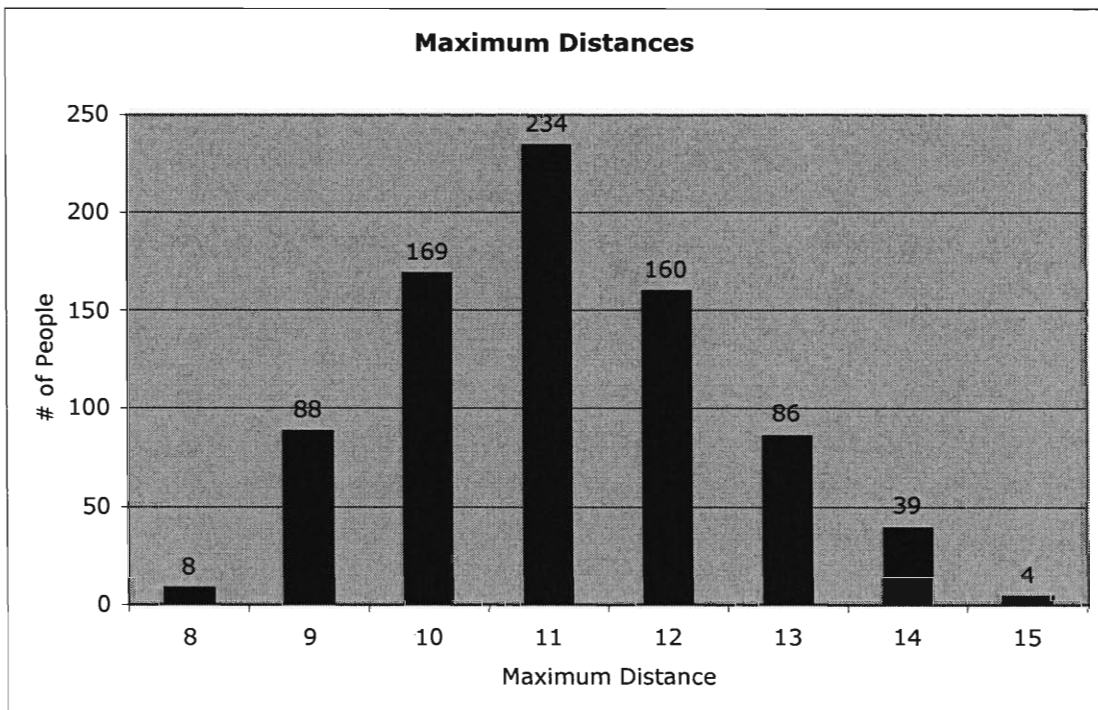


Figure 5-1: Distribution of AIMBase Maximum Distances

The horizontal axis of Figure 5-1 shows the maximum number of steps from each person to any other AIMBase participants. The vertical axis shows the number of users in each category. For example, there are 88 people who can reach everyone else in the graph in a maximum of 9 steps, and the majority of people (nearly 84%) can reach everyone else in only 12 steps.

6 – Survey Results and Analysis

6.1 – Geographical Clustering

One of our strongest initial expectations was that geographical clustering would be high. In our case, this means that people who know each other over AIM would generally live in close proximity to each other. This follows on the heels of our hypothesis that most users of AIM would be college students, which was based almost entirely upon our personal experiences with the network.

Our survey results showed that 75% of all respondents to the survey described themselves as students born after 1984, which generally means that they are college students, barring an extraordinary number of students who have been held back. Even allowing for small variations in the year students entered college, it is still safe to say that more than 70% of all survey respondents are college students.

It is not necessarily intuitive that high numbers of college students should lead to high geographical clustering. Consider, however, that most college students are highly likely to know other people at their college. In addition, most colleges provide continuous Internet connectivity, which means that students are likely to use AIM and leave it running continuously in order to maintain both synchronous and asynchronous communication with their friends.

In order to pull useful information out of our survey results, we purchased a commercial database that allowed us to translate U.S. zip codes (the location data collected by our survey) into latitude and longitude coordinates. At the time this report was written, only 2 of 297 survey responses listed locations outside the U.S. (one in

Germany, and one in Switzerland), so we felt that using only U.S. locations generated sufficiently accurate measures of average distances and similar metrics.

We then used a publicly available set of software called PROJ4, provided by the U.S. Geological Survey, to calculate accurate distances between latitude and longitude coordinates. Since we collected both home zip codes and work/college zip codes, we decided to build three distance sets. Each distance set includes a list of all the pairs of users who have indicated they know each other, plus the distance of that particular link. For example, if a user AimGuy1 in Boston has indicated that they know a user AimGirl2 in Worcester, the distance set will include the row “AimGuy1, AimGirl2, 60.5”, where the number is the distance, in miles, from AimGuy1’s location to AimGirl2’s location.

The first distance set contained all the distances using users’ home zip codes as their locations. So, following on the example above, even if AimGuy1 and AimGirl2 went to college in Miami and Boston, respectively, they would still be measured as 60.5 miles apart since they claimed Boston and Worcester, respectively, as home addresses.

The second distance set measured distances based on college zip codes. This distance set is very similar to the previous one, except that home addresses are ignored in favor of college addresses. It should be noted that the number of survey responders with college zip codes differs from the number of users with home zip codes, although the same statistical methods can be applied.

The third distance set measured the minimum link distance between two users. In the example above, the distance between AimGuy1 and AimGirl2 would be 60.5 miles. The home addresses are much closer than the college addresses. We chose not to measure college-to-home link distances and made the assumption that college students do not

meet very many people outside of the college they attend (or other colleges with which they closely interact). They may know the faces of some people from the city they go to college in, but they will rarely converse with them over AIM. We conducted a casual survey of several college students chosen randomly from the database, and only two of the fifteen people we talked to know anyone at all from their college zip code who was not also a college student in that area.

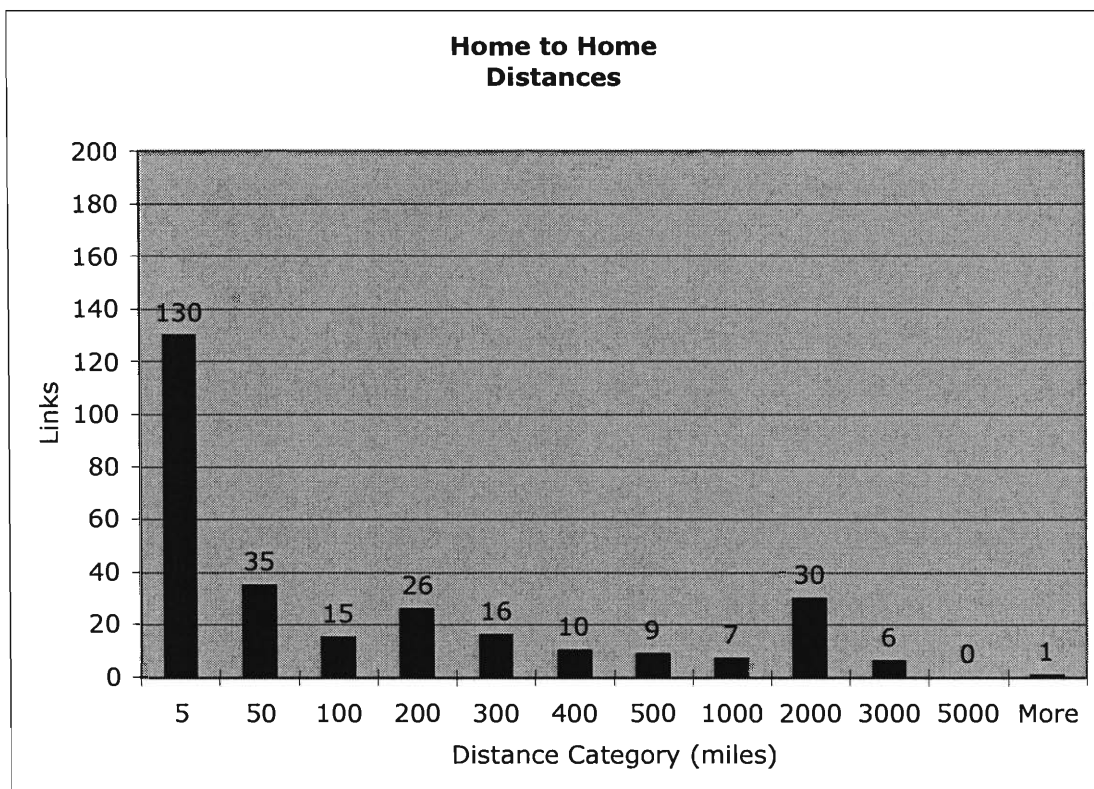


Figure 6-1: Distribution of Link Distances (Home-to-Home)

Figure 6-1 is a histogram of the first distance set, measuring link distances between home addresses. The vertical axis represents the number of links in a category, and the horizontal axis represents the upper bound on a distance. For example, 130 links traverse 5 or fewer miles, while only 1 link crosses more than 5000 miles. To some

extent, this data proves our hypothesis is true; the more links to the left of the graph, the higher the geographical clustering, and there is a definite trend towards shorter links. On the other hand, only 45% of the links are within 5 miles of each other, and only 57% are within 50 miles, so it seems at first glance that a not-insignificant percentage of links occur across large geographical distances.

As we mentioned earlier, a large percentage of survey respondents are college students. Many college students go away for college, where they meet new friends, who they may begin to talk with over AIM. A graph of the second distance set, measuring distances between college zip codes, shows a more promising distribution.

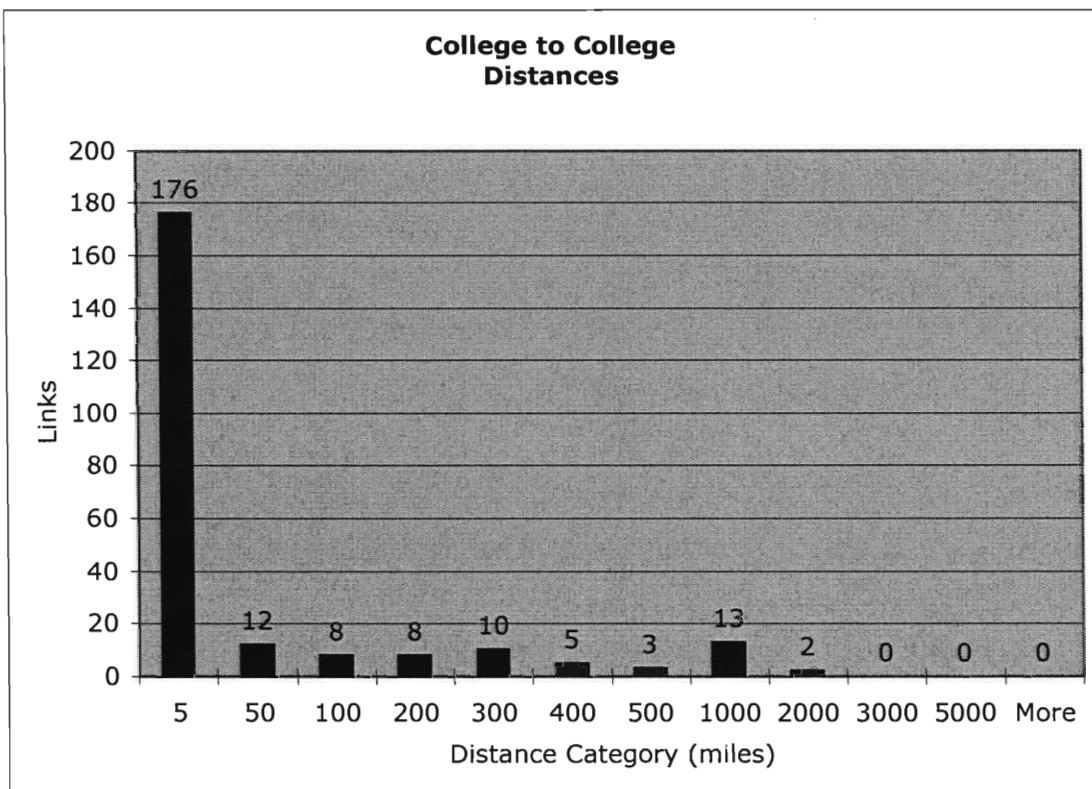


Figure 6-2: Distribution of Link Distances (College-to-College)

The histogram in Figure 6-2 shows that a much greater percentage of the links traverse short distances. With 74% of the links reaching less than five miles, and 79% of the links within 50 miles, it seems that our hypotheses may indeed be as accurate as we had hoped. There are 48 fewer links in this histogram, because not all respondents go to college or work away from home.

After creating these graphs, we realized that we were not seeing the whole picture. While some people meet other people at college, high school students, family members, and others will know each other from their home address, so we combined the two previous distance sets and selected the minimum distance; either home-to-home or college-to-college. First, we selected only those people who responded to the database who had both a home address and a college address, and then we generated the same type of data set that generated the previous two charts and discovered that the encouraging results presented by the second data set were even more obvious.

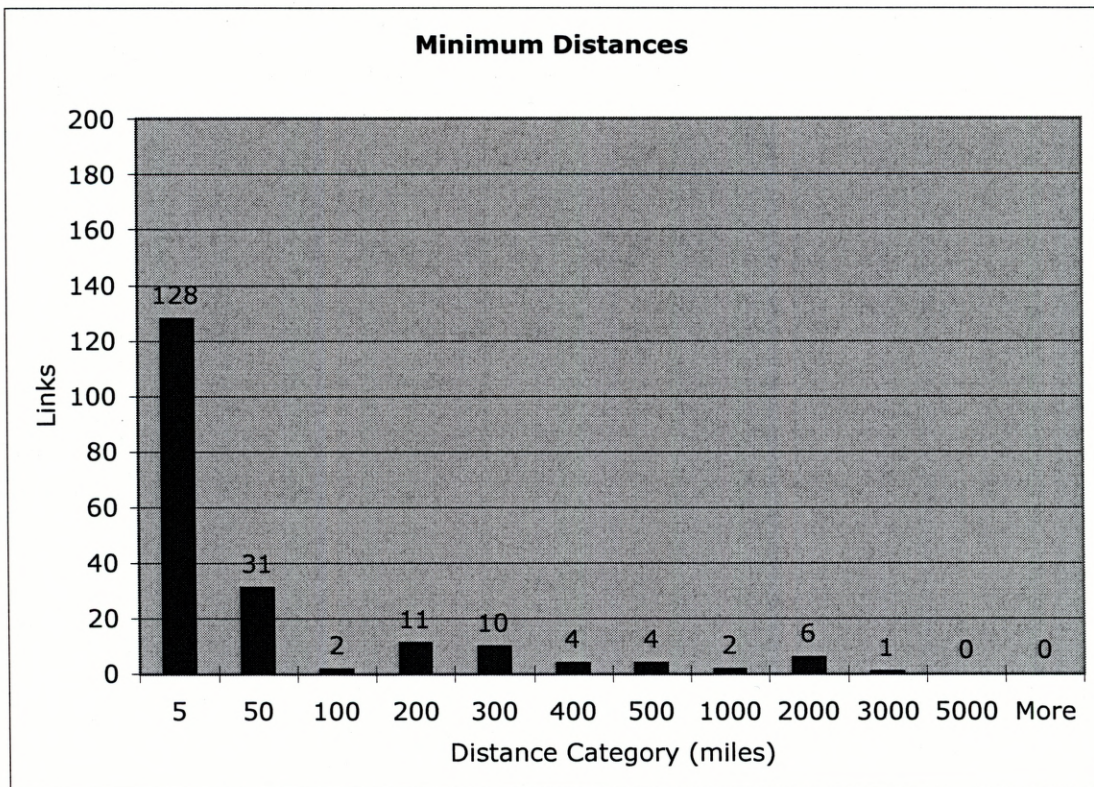


Figure 6-3: Distribution of Link Distances (Minimum)

In Figure 6-3, there are only 200 links shown, because there were only 200 respondents to the survey that fit the criteria that they have both home and college zip codes. Again, the large majority of links (64%) traverse very short distances, and again, 79% of the links fall within the 50-mile boundary.

We have concluded, based upon the charts above and further study of the distance sets, that geographical clustering within the AIM network is extremely high. Obviously, this has some implications for how AIM clients are used.

6.2 – AIM and Face-to-Face Interaction

One of the obvious implications of geographical clustering is that the number of people who interact with their AIM contacts in person (face-to-face) should be very high. Our survey collected both the number of people in a person's buddy list, and the number of those people that they knew in person.

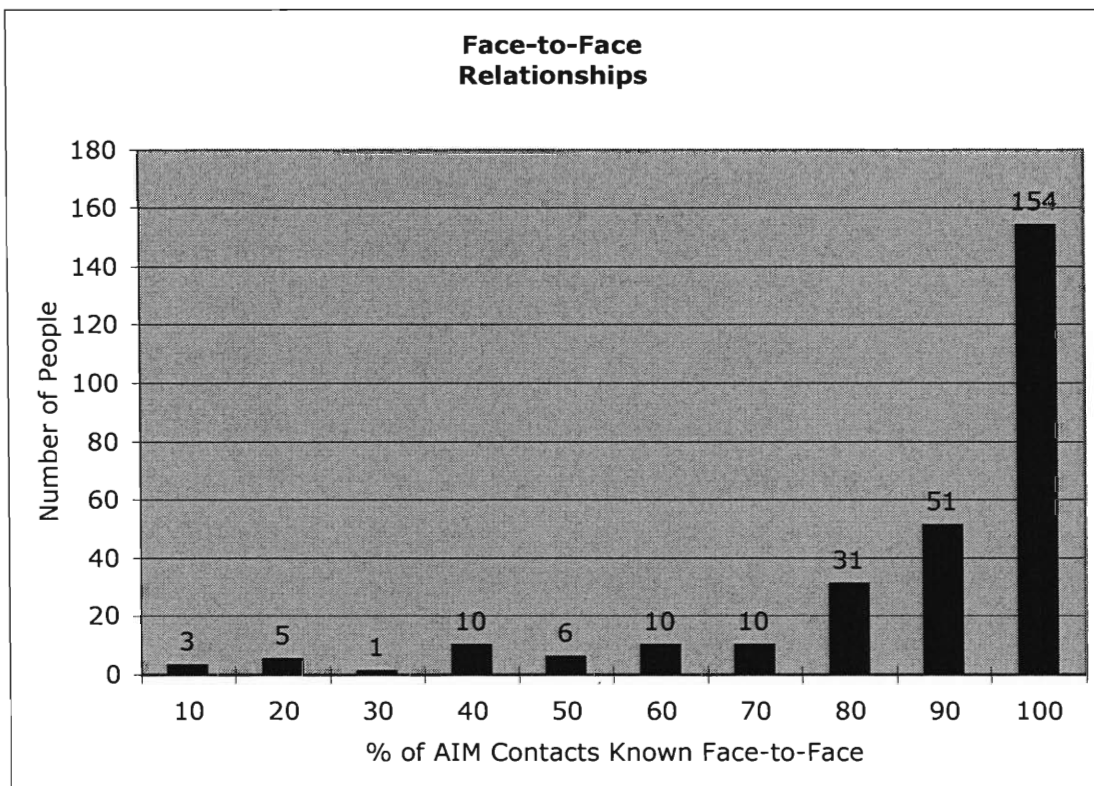


Figure 6-4: Online vs. Face-to-Face Relationships

The horizontal axis of Figure 6-4 shows the percentage of people in any given person's buddy list, or list of aim contacts. The vertical axis shows the number of people who fall into each percentile range. For example, 3 people know 10% or fewer of their AIM contacts face-to-face, but 205 people know 90% or more of their AIM contact in person.

Notwithstanding our expectation, reinforced by the high geographical clustering, we were not expecting such a close connection between relationships over AIM and “real life” relationships. One explanation is that, while AIM does have a feature that lets users search for other people with similar interests, very few people actually fill out the information to enable this feature. In other instant-messaging environments, notably ICQ, this feature is one of the main reasons for the formation of quick connections across large distances.

Connections that span large social “distances”, as from a random search for people with vaguely similar interests, are generally called “random links” in social network theory. Random links are the reason that small-world networks form. If everyone only knew people near them, it would take a very long time for a message to get from Los Angeles to Boston, but Milgram and others have shown this to not be the case. One way to think about how random links increase the closeness of people in a network is to consider the lattice model.

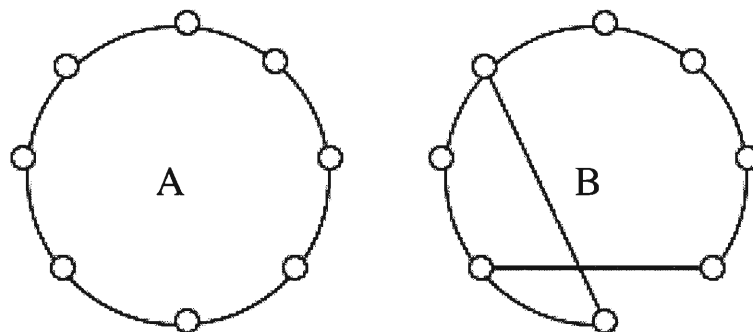


Figure 6-5: The Lattice Model

In Figure 6-5, the two large circles A and B are examples of lattices. Each small circle is a node, and each line segment is a connection. Now, imagine that instead of

abstract nodes, each small circle is a person. Each person can talk to the people they are directly connected to, and people one step past that, so, in lattice A, each person can talk to 4 other people. If the person at the bottom of lattice A wants to get a message to the person at the top of the circle, there must be at least one intermediary. If this lattice had many more people, it would take proportionally more steps.

On the other hand, lattice B has had a few random links added. These are similar to the geography- or society-spanning links that occur in social networks (such as when ICQ members befriend people they find through the search feature). Note that now the person at the bottom of the lattice can talk to the person at the top of the lattice. As more and more normal links are replaced with random links, the average distance between people on the graph decreases (at least up to a point, eventually, the randomness makes everyone just as far apart as in the original lattice).

Since AIM users usually know each other in person, it stands to reason that the social network measures that apply to real world interaction will be very similar to those that apply to AIM. On the other hand, we feel that it would be worthwhile to study a network such as ICQ where random connections occur more frequently.

6.3 – The Effect of Instant Messaging on Phone and Email Usage

Before the advent of AIM, telephone conversations and email generally filled the “quick message” niche. It was our impression that AIM would have significantly affect phone and email usage. Instant messaging clients are generally unobtrusive and extremely easy to use. Unlike telephone conversations, AIM conversations can be easily carried on without devoting full attention to the process. Responses can be delayed as

much or as little as one desires, and there is no fee. Unlike email, AIM can be fully synchronous, where both parties can talk simultaneously and respond to each other quickly, whereas email requires users to specifically check for new messages. In addition, while AIM conversations can be archived, the official AOL client does not support this, so archives are uncommon. Nearly all email clients support the saving of messages, and most people save their email communication. Since most people expect AIM messages to be ephemeral and email to provide a record, the two mediums can be used for different and distinct tasks.

Our conjectures were apparently accurate, as can be seen in Figure 6-6 and Figure 6-7. 70% of AIMBase survey respondents felt that their email usage had decreased as a result of AIM, and 79% of respondents felt that their phone usage had decreased. Email usage has likely decreased less than phone usage because instant messaging more closely mirrors the synchronous nature of telephone conversations than it does the asynchronous nature of emails. While AIM can be used asynchronously, the limit on message sizes, formatting, and formality available in email discourages its use as a replacement for email.

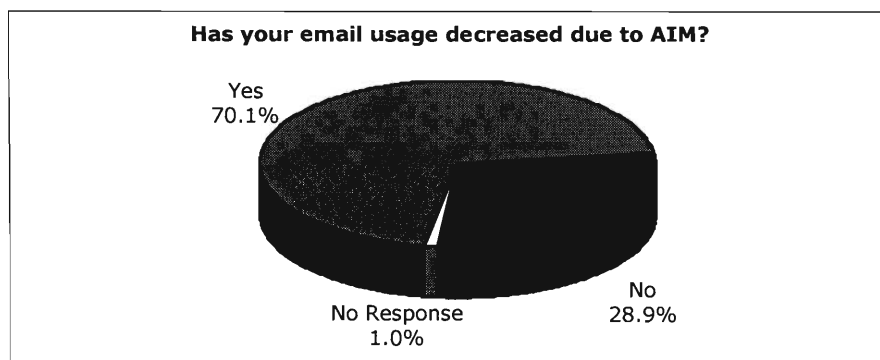


Figure 6-6: Changes in Email Usage Due to AIM

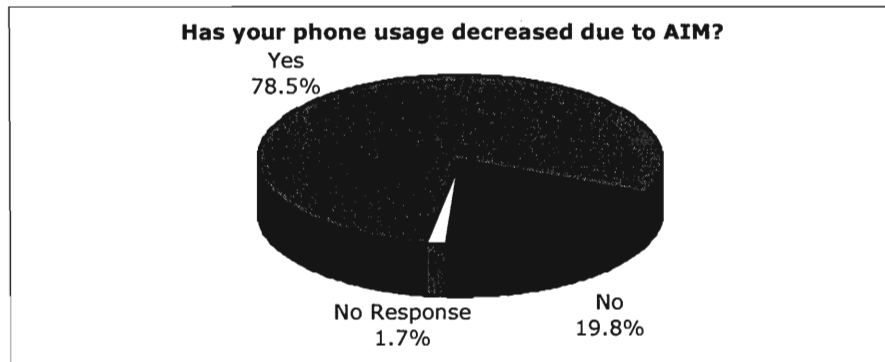


Figure 6-7: Changes in Phone Usage Due to AIM

6.4 – Information Propagation

One of the notable differences between the AIMBase study and Milgram’s Small World research is that his study required people to guess the next closest person to the target. This is not necessarily easy; most people don’t know who the next closest person to the president is. In our study, instead of asking people to get a message somewhere, the message is being propagated in a manner akin to a virus. This means that people do not need to know how they are connected to someone six degrees away in order for us to collect that data.

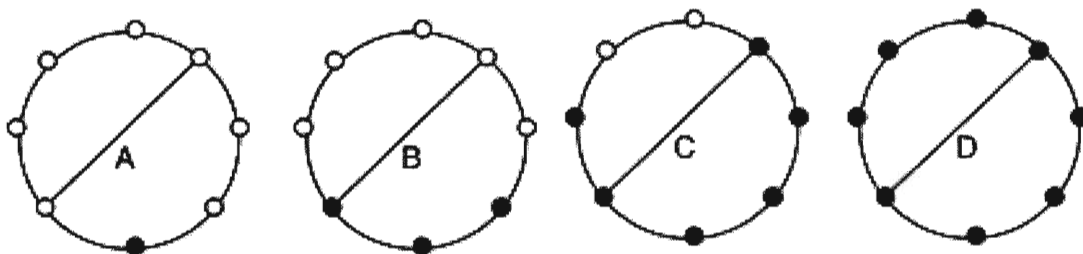


Figure 6-8: Disease Front Propagation

One of the things we wanted to study was the way in which information propagated over the AIM network and how that compares to the spread of disease fronts. (As in

Figure 6-8.) One of the primary indicators of disease epidemics and other broadcast propagations is that they fit into a log-normal curve (or a series of log-normal curves), with a steep upslope and a gradual downslope. In general, the steep upslope indicates that a carrier of the information or disease comes in contact with many uninformed or uninfected people. (Gordis 365)

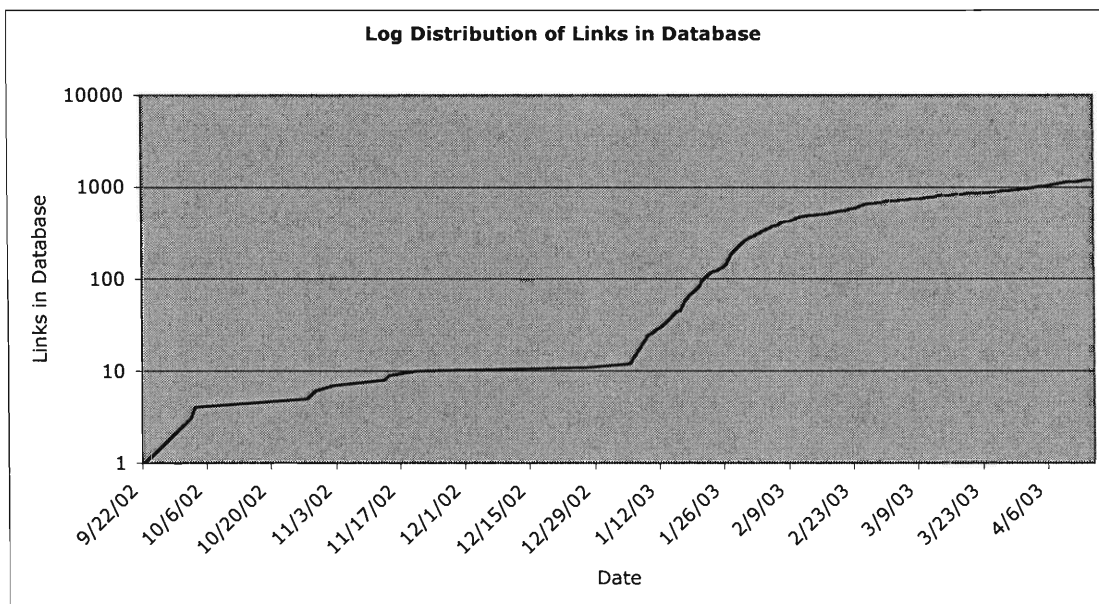


Figure 6-9: Log Distribution of Links in Database

Figure 6-9 shows a logarithmic graph of the number of links in the database from September 22, 2002 through April 15, 2003. There are in fact two separate log-normal curves, one at the beginning of the graph, and one at the turn of the year (around the middle of the graph). In epidemiology, each sharp upslope indicates a place where an infected person comes in contact with a large number of uninfected people (for example, when an infected person gets on a plane and travels to another country). In our case, the jumps indicate that a hub person adds the AIMBase link to their profile. Note that since the graph in Figure 6-9 is plotted on a logarithmic scale, each jump is different in size;

the first one may be a similar percentage increase in the number of links in the database, but it is a much smaller increase in absolute terms.

Another way of discovering hub users in the database is by noticing when large bursts of links are added. By tracking the number of links added per day, we can discover when new people add the AIMBase link to their profile, which allows us to discover all of the hubs, and not just the ones that are spectacular against the general size of the database.

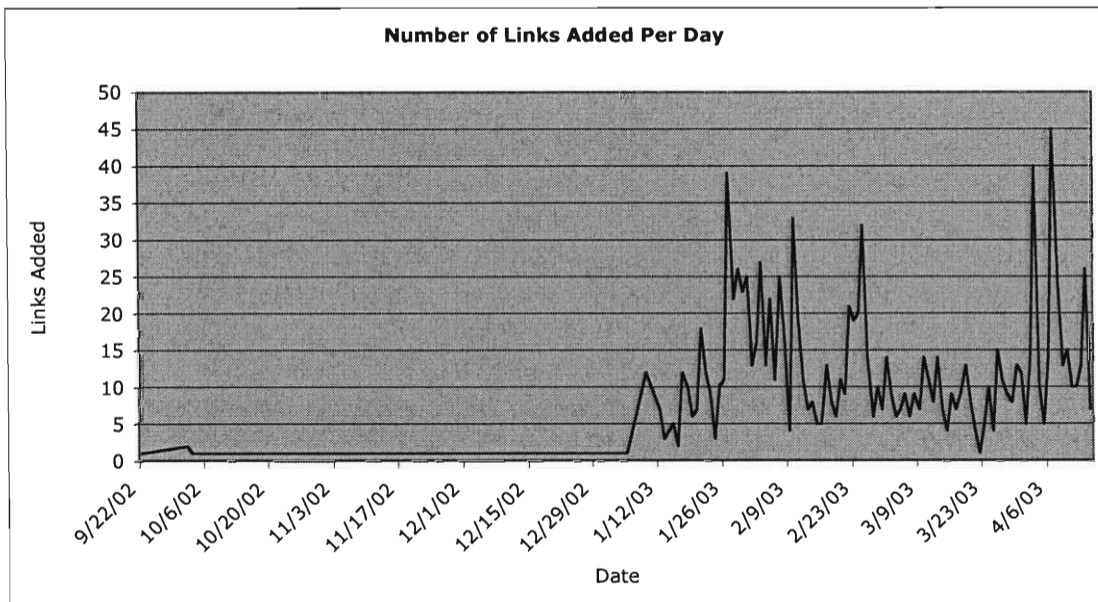


Figure 6-10: Links Added Per Day (Hub Discovery)

Most of the large spikes in Figure 6-10 represent days when particularly popular users began participating in the AIMBase project (by adding the AIMBase link to their profile). For example, on April 7, 2003, 60 new links were added, with only 6 participants as the people with the most activity.

Even though the high clustering of the AIM network means that people are not necessarily as closely connected as we had expected, the epidemic nature of information

transmission across the network means that ideas and memes can spread very quickly. We think that the quick spread of “wildfire” phenomena on the Internet is due in part to the epidemic nature of the AIM network and other instant messaging networks.

6.5 – Instant Messaging Productivity

Workplace collaboration software, often with features that allow instant communication, has existed since it became affordable to install networks. Oddly enough, most businesses still see the popular instant messaging networks as purely social tools. Many of the larger instant messaging providers are now trying to change the prevailing attitude in order to make money from what is mostly being provided for free. Advertising for the service provider (such as AOL) and banner ads only go so far towards offsetting the cost of instant messaging networks.

In order to expand the network into the corporate sector, and bring in more revenue, instant-messaging providers, including Microsoft, AOL, and the open-source Jabber project are designing “enterprise” IM solutions. In many cases, they modify their clients slightly to focus on productivity, and provide special administration tools to regulate IM usage. Unlike traditional collaboration tools, these enterprise IM tools interoperate with the larger public networks. Office managers can generally block interaction with the public network, but the opportunity still exists for widespread IM interaction.

One section of our survey dealt with users’ impressions of how they used AIM in workplace and academic environments. The first two questions asked whether AIM was

used at all for work or academic purposes, and the third question asked the user's impression of AIM's effect on their productivity.

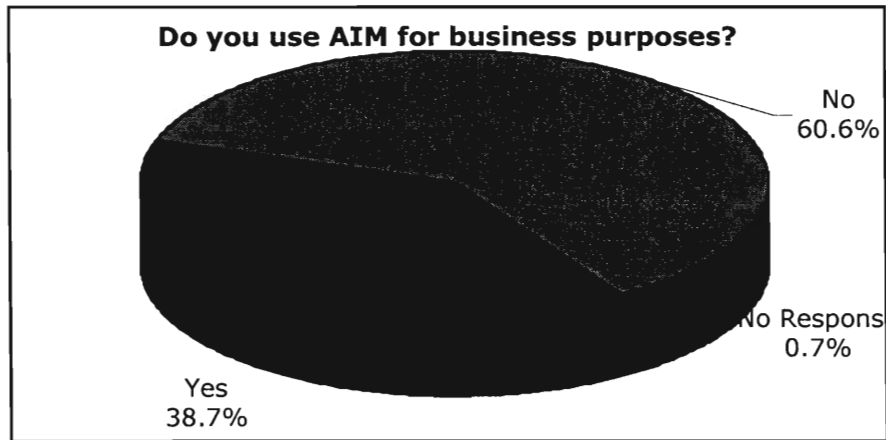


Figure 6-11: Survey Respondents Using AIM for Business Purposes

Only 39% of the people who took they survey responded that they used AIM for business purposes. We believe that that is actually a very high percentage, since so many of the survey respondents were, as can be seen, students. The 39% measure is of all the people who took the survey, not just those with jobs.

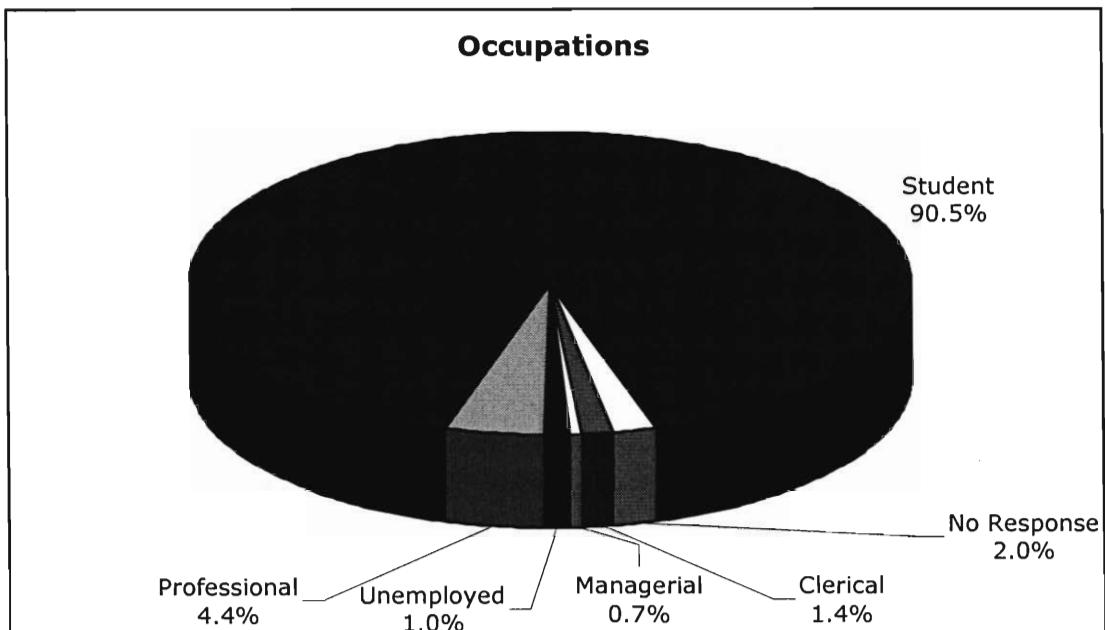


Figure 6-12: AIMBase Participant Occupations

When we look at academic usage, we get a much larger number of positive responses, which is to be expected with the user distribution we have discovered. As can be seen in Figure 6-13, 88% of survey respondents indicated using AIM for academic purposes, which is nearly the percentage of survey respondents who are students. This seems to indicate that AIM becomes very much a part of a student's lifestyle.

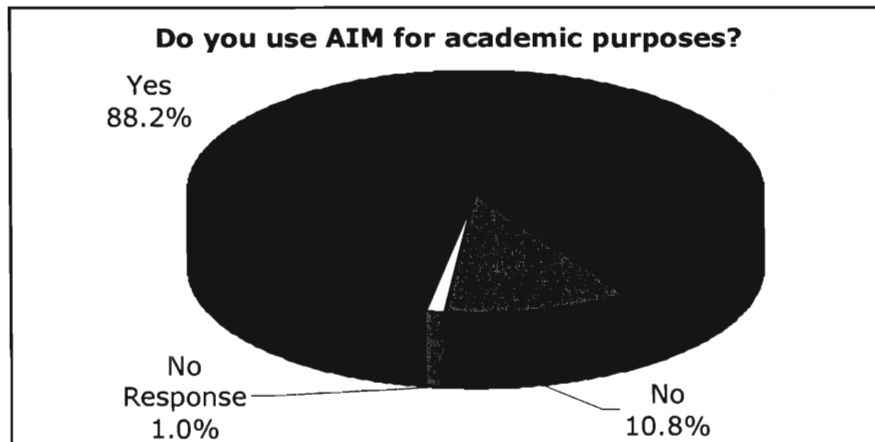


Figure 6-13: Survey Respondents Using AIM for Academic Purposes

Interestingly, very few AIMBase participants felt that AIM had boosted their productivity. In fact, almost as many people felt AIM had decreased their productivity. We expect that this is due generally to the distracting nature of instant messaging clients. If a college student is trying to do homework, they can close the door and not check their email, but AIM users will generally not sign offline. In our experience, many people will talk with their friends via the AIM network even when they recognize that they are procrastinating. Because of the immediate nature of instant messaging, it hard to ignore messages from other people, just as it is hard to ignore a phone ringing or someone tapping you on your shoulder.

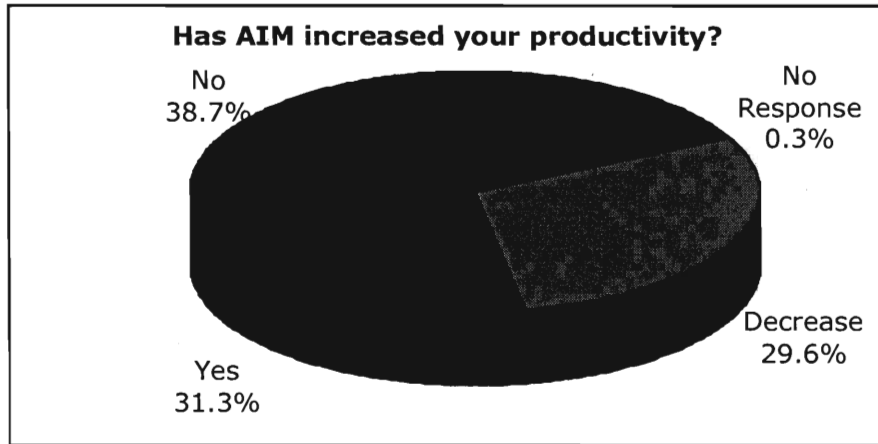


Figure 6-14: Survey Respondents Using AIM for Increased Productivity

6.6 – Demographics of AIM Users

In order to compare the AIM-user demographic to the general online population, we collected age, gender, location, and occupation information as part of our survey. Previously, we have looked at geographic distance as it relates to clustering of AIM users, and occupation as it relates to productivity, but this information is also interesting in comparison to statistics collected by other surveys.

6.6.1 – Age

Harris Interactive, conductors of *The Harris Poll*, frequently runs polls on Internet usage. In February 2003, they released a public article containing age and gender information for the general adult online population. A chart showing the general distribution of online adults is shown in Figure 6-15. As can be seen, Internet users are evenly distributed among all but the oldest age set.

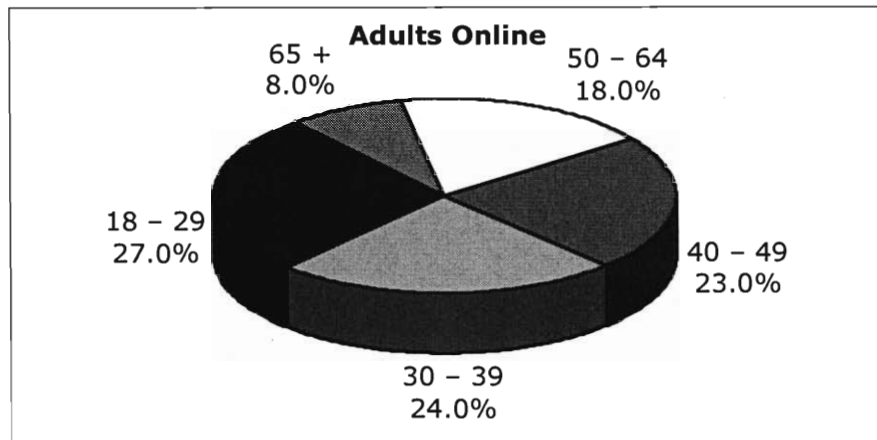


Figure 6-15: Harris Poll of Ages of Adults Online

Unsurprisingly, AIMBase participants do not match this egalitarian distribution. Instead, the vast majority of AIM users seem to be college age students, with the second largest age group being even younger. Unfortunately, the Harris Poll data does not include information about people under 18. As a large portion of AIM users are in high school, we decided to use information from all age brackets. (All personal identifying information for submissions indicating an age under 13 was removed, in keeping with online survey regulations and etiquette.) (Taylor)

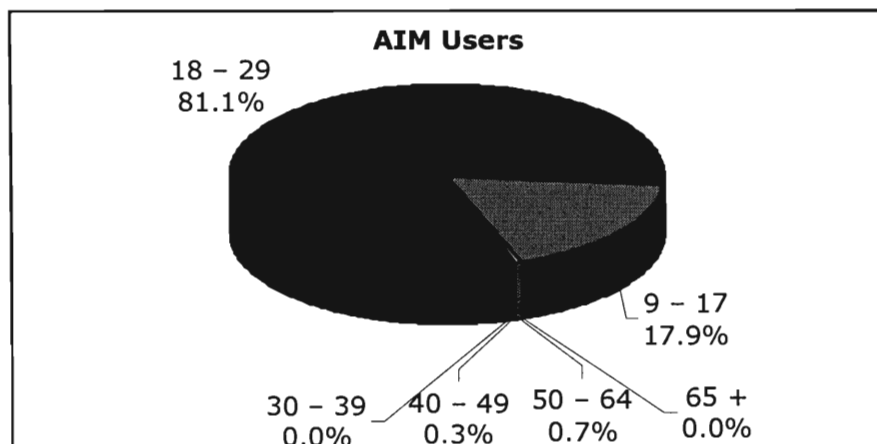


Figure 6-16: Ages of AIMBase Survey Respondents

Admittedly, survey results for AIMBase may be somewhat skewed. Although the survey itself is easy to take, and joining the AIMBase project was made as simple as possible, many older people are extremely cautious about divulging information online. Highly popularized cases of identity theft and online fraud have made it nearly impossible to collect completely accurate demographics online. Generally, younger people with more online experience and understanding of the risks are more likely to take surveys, which is likely reflected in our data about age and occupation. In addition, our initial hubs were college students, so the spread of connections started with a small selection of biased hubs.

6.6.2 – Gender

Gender, on the other hand, is less likely to affect survey responses, and shows a not unexpected trend. According to *The Harris Poll*, Internet users are just as likely to be male or female (a 50/50 split). According to our data, it is more likely that AIM users will be male. There are several possible explanations; none of them are likely the whole story. Our primary hypothesis is based on data that males spend more time online than women (24% more, according to a 2002 Nielsen survey). More time online means that AIM is a more likely method for interaction (in lieu of phone or face-to-face meetings).

(Cyberatlas)

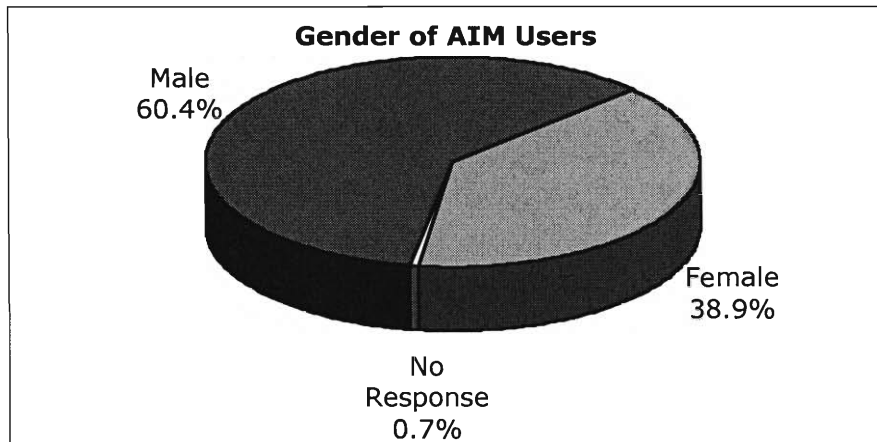


Figure 6-17: Gender Divisions of AIMBase Participants

6.6.3 – Geography

With such high geographical clustering, we were half-expecting all of the zip codes to be clustered almost entirely in Massachusetts. We were pleased to find that there was a fairly broad distribution across much of the U.S.

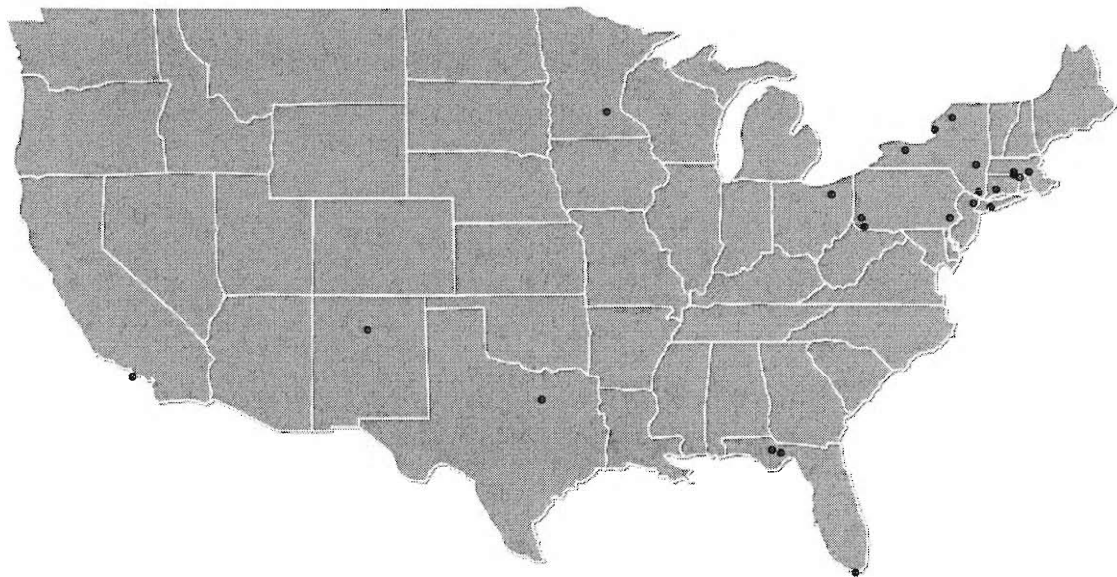


Figure 6-18: Geographical Distribution of AIMBase Participants' Home Addresses

In fact, while many users are clustered in New England, this can partly be attributed to the high density of colleges and people in the area. Likewise, the dearth of people in the Midwest can be attributed to the sparsely populated nature of the region.

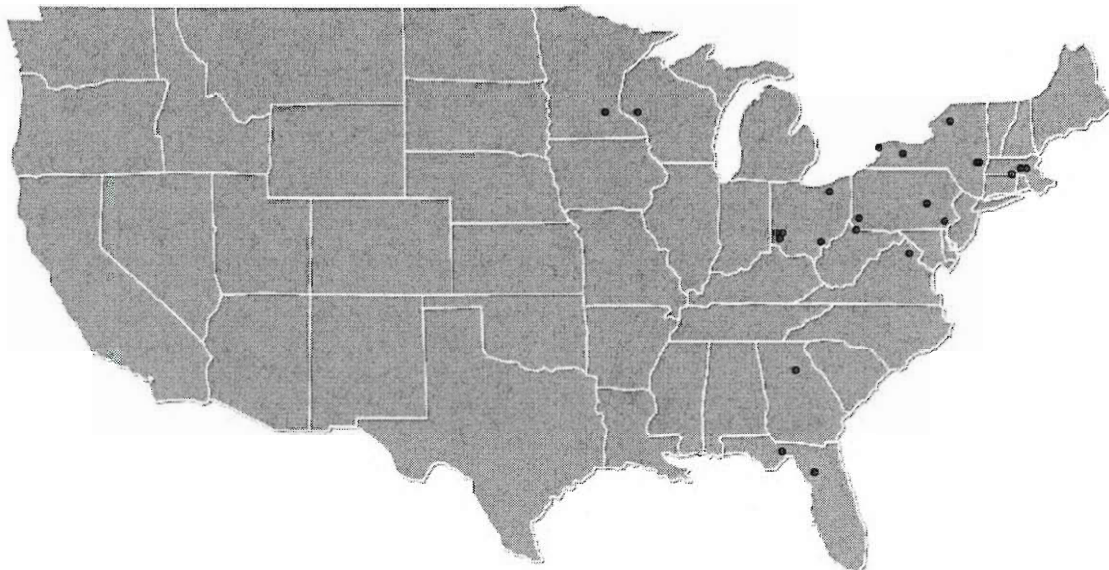


Figure 6-19: Geographical Distribution of AIMBase Participants' College Addresses

When we look at the college addresses, we see a larger distribution along the east coast. As we continue to run the project and survey after this report is complete, we expect that the distribution will spread more as hub users in other regions begin to spread AIMBase.

6.7 – Cliques and Demographics

In Section 4.4 – Finding Cliques, we described how we found cliques based on a threshold number of people that each person in the clique must know. When run with a threshold value of 7, the heuristic returned a complete 7-clique. A clique is a completely

connected graph or subgraph. The phrase “7-clique” means that every member of a group of 7 was connected to every other person in that group.

When run with a threshold of 6, it returned the same 7-clique, and a dense subgraph consisting of a group of WPI students.

When run with thresholds of 5 or lower, the program returned a connected subset containing many possible dense subgraphs. The difficulty was in separating these subgraphs. In the previous case, the two subgraphs were not connected, allowing a simple breadth-first search to separate them. However, in this case, many of the subgraphs were connected by one or two links. In order to separate these connected subgraphs, we designed another software tool based on a clique-finding heuristic. The program started from each user in the connected set of dense subgraphs, and tried to build a clique around them. It did this by selecting all users in the subgraph that were within a maximum of two steps. The program then printed the members of each user’s clique. Unfortunately, this requires manual operation, and returns many similar subgraphs. A better automated solution to the problem of separating dense subgraphs is a possible area for further research.

The dense subgraphs we found turned out to imply a very strong correlation between the dense subgraphs of AIMBase participants and their social connections.

One of the dense subgraphs contained 22 people who are all closely involved with the WPI community, where this project was conducted. In fact, it contains the two project administrators. All of these users gave their “away” zip code as 01609, the WPI zip code, although their home zip codes range from Florida to New Hampshire. All of the users in

this group answered that they used AIM for academic purposes. 83% percent of them said that their telephone usage decreased due to AIM.

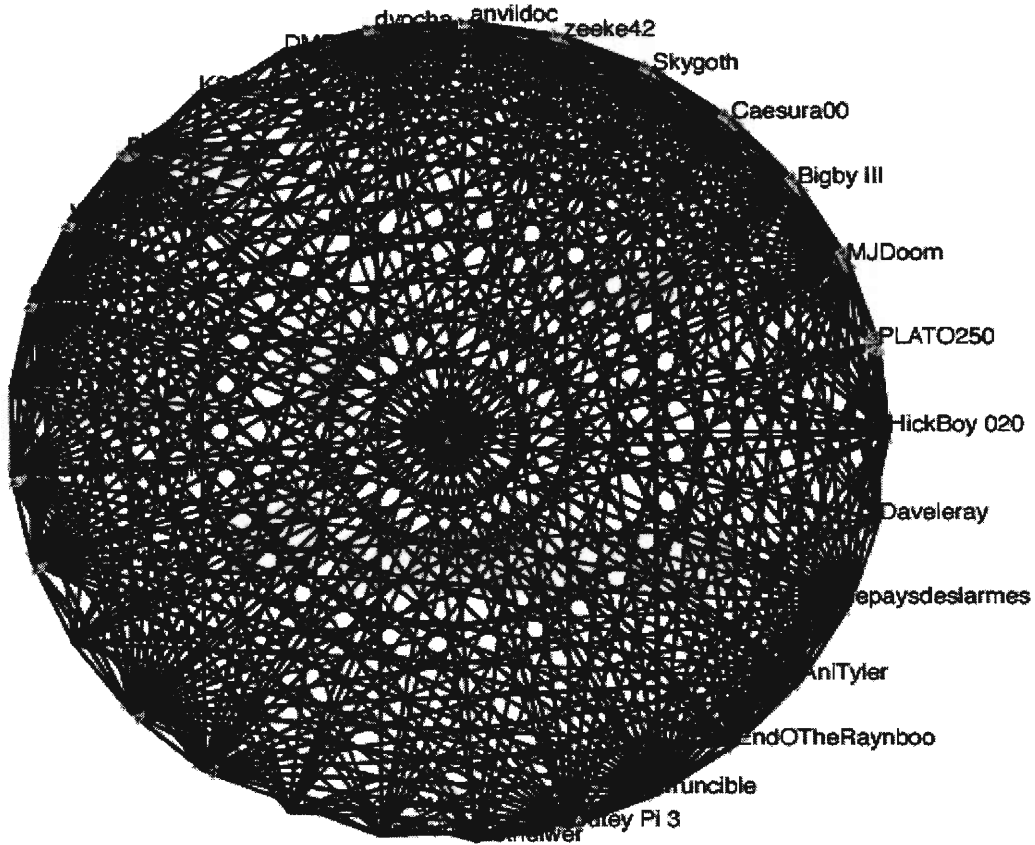


Figure 6-20: WPI Clique (Dense Subgraph #1)

Another dense subgraph, shown in Figure 6-21, consists of only 7 people, but has some interesting properties. Unfortunately, we had only one survey response from the group. This user answered that he was a college student in Michigan who used AIM for academic purposes. We conjecture that this is another group of college students, but we have no way to know for sure. Interestingly this sub-graph is a true clique, a complete graph of seven nodes. That suggests that rather than just being a group of friends, this is an organized group of some kind. Without organization, the graph would probably not be complete, because not everyone would have everyone else's name. If it were an

organized group of some kind, with a list of AIM usernames distributed to all members, then the graph would be complete.

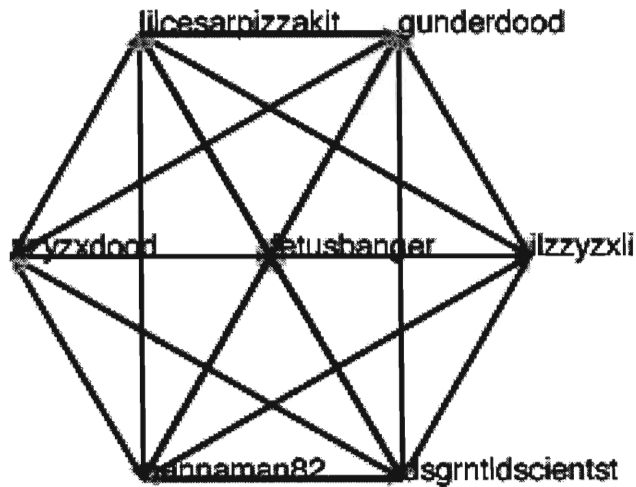


Figure 6-21: 7-Clique (Dense Subgraph #2)

A third dense subgraph contained 11 people, all of whom listed themselves as being from the same geographic area. The home zip codes for all of these users began 323 except for 33912. These zip codes are in the Tallahassee, Florida area. The birth years were all within the range of 1980 to 1985 with the majority between 82 and 84. Everyone in this group is a student away at college. Their college locations range from nearby in Florida, to Massachusetts, to Indiana. This data suggests that this is a group of high school friends who use AIM to keep in touch while away at college. This is borne out by the fact that this clique contains one of the project administrators and he confirms that several of the members are high school friends.

A fourth group also emphasized the relationship between the density of connections in the AIMBase graph and geographical proximity. Each of the seven people in this group who answered the survey goes to college in Pittsburgh, PA. They all say that

they use AIM for academic purposes, but interestingly, 75% of them also say that it has decreased their productivity.

The final group discovered by our process for finding dense subgraphs consists of 16 people who all attend college in Dayton, Ohio, with the exception of the two project administrators. The reason for our inclusion in this clique is likely due to the discovery of the project through one of our (publicly displayed on the website) AIM profiles.

Everyone in this group said that they used AIM for academic purposes. The reports of its effect on productivity with this group were mixed. Also notably, this group is almost entirely male.

7 – Technology

7.1 – Web Server

The project web server runs Apache 1.3.26 httpd software. One of the challenges in implementing the system was the fact that the web browser built into AIM does not completely adhere to the HTTP specification. Space characters in the request must be converted to the escape sequence '%20'. The AIM browser does not do this. In order to work around this problem, we wrote an Apache mod_perl module to intercept the request on its way through the server and rewrite it with the proper escapes. Mod_perl is an Apache component that allows user written software to perform various operations at nearly any point in the request processing cycle. Upon moving from our test bed server to our production server running the latest version of Apache, this code ceased to work. We discovered that the apache software had been updated to discard these malformed requests earlier in its processing pipeline. We had to modify the Apache source code to pass these requests far enough for our program to fix them.

8 – Conclusions

Although we ran into both time constraints and technical trouble (especially finding dense subgraphs) we managed to collect a fair amount of useful information. First and foremost, we discovered that geographical clustering greatly affects how AIM communities form. This likely means that the AIM network reflects, albeit imperfectly, the social networks of people who know each other in person.

We also discovered that it is possible to find communities of users with similar demographics based on their placement within a network. This is more surprising than it seems at first glance: generally it is easy to start with a social group and show how they connect as part of a graph, but it is often difficult to go the other way and use a graph to find social groupings.

As an example, consider the network of actors used by the Oracle of Bacon. If we knew which movies every fifth actor had been in, we could probably find the same information for the other 80% of actors based on their place within the network. On the other hand, we probably couldn't start with what kind of coffee 20% of the actors liked and determine the coffee preferences of the others. What is so interesting about the AIMBase graph is that we can start with a small sampling of data and determine the demographics of most of the users. For instance, if we know that 3 people in a dense subgraph of 10 are college students, we can say with high confidence that the other 7 will also be college students. We have found that the same is true for geographical location, occupation, and even gender. From an applications perspective, this has strong implications. For example, if AOL wanted to try a targeted marketing campaign on its users, they could survey only a small subset of users, then use the graph of AIM users to

guess the demographics or tastes of the rest of the users. As instant messaging becomes even more popular, data mining based on the graph of IM networks could provide a useful alternative to mass surveys or other current methods.

8.1 – Sample Size Disclaimer

According to the generic calculation of appropriate sample size, 384 people chosen completely at random out of a population of 140 million will produce accurate results with 95% confidence. This is based on the standard power curve for determining sample size. Unfortunately, because of the broadcast-style propagation of the AIMBase project, our sample is not very random. The initial broadcast points were college students, and, although people in other occupations have participated in the experiment, we expect that our survey results are not a representation of the AIM network as a whole, but rather of the particular groups that are connected to college students and find the study of enough interest to participate.

8.2 – Recommendations for Further Research

Although we intend to continue collecting data after the finalization of this report, there are plenty of issues that our current experiment does not consider. One of these is the aforementioned comparison of groups within the AIM network to interest groups in real life. We expect that the correlation will be high, but there may be interesting

contrasts, where interest groups over AIM are more likely to interconnect than ones in real life.

Branching out into other IM networks may provide useful data as well. The ICQ network, with its easy discovery of random users, may actually form a better “small world” than the AIM network. Microsoft’s new peer-to-peer application “threedegrees”, or “3^o” may be an interesting medium to study; unlike generic instant messaging clients, it simultaneously connects one person to a group so that almost every interaction is broadcast to all members of a group.

With further analysis, it should also be possible to make useful generalizations about the propagation of information via the Internet based on the spread of the AIMBase project. The growth data gathered tends to imply that a small number of people are the primary information distributors for the AIM network, and it would be interesting to see if the same is true for the Internet as a whole.

9 – Additional Information

Additional information is available on the CD included with this report, including the database of connections and surveys, as well as the archived code for the website, the AIM ‘bot, and various graph-traversal tools.

10 – References

America Online Announces Expansion of AOL Amber Alerts Service to Millions of AIM

Users and Internet Users. AOL, Inc Press Releases.

<http://media.aoltime Warner.com/media/newmedia/cb_press_view.cfm?release_num=55252884> (March 19, 2003)

America Online: Who We Are: Fast Facts. AOL, Inc Press Releases.

<<http://corp.aol.com/whoweare/fastfacts.html>> (March 19, 2003)

Cyberatlas staff. *Men Still Dominate Worldwide Internet Use.*

<http://cyberatlas.internet.com/big_picture/demographics/article/0,,5901_959421,00.html> (April 22, 2003)

Greengard, Michael. *Science/Six Degrees Experiment.*

<http://www.urbanlegends.com/science/six_degrees_experiment.html> (April 30, 1997)

Grossman, Jerry, et al. *The Erdős Number Project.*

<<http://www.oakland.edu/~grossman/erdoshp.html>> (March 15, 2003)

Taylor, Humphrey. *Those with Internet Access to Continue to Grow but at a Slower Rate.*

<http://www.harrisinteractive.com/harris_poll/index.asp?PID=356> (April 22, 2003)

Lee, E-J., C. Nass, *Experimental Tests of Normative Group Influence and Representation*

Effects in Computer-Mediated Communication. Human Communication Research, July 2002.

Milgram, Stanley. *The Small World Problem.* Psychology Today, April 1967.

Gordis, L. *Epidemiology Second Edition.* Saunders, June 15, 2000.

- Postmes, S. T. Russel, M. Lea. *The Formation of Group Norms in Computer-Mediated Communication*. Human Communication Research, July 2000.
- Reynolds, Patrick. *The Oracle of Bacon at Virginia*. Dept. of Computer Science, University of Virginia. <<http://www.cs.virginia.edu/oracle/>> (March 26, 2003)
- Roberts, B., D. Hendrickson, *Six Degrees of WPI*. Worcester Polytechnic Institute. <[http://users.wpi.edu/~flicker/IQP/Six Degrees of WPI.pdf](http://users.wpi.edu/~flicker/IQP/Six%20Degrees%20of%20WPI.pdf)> (October 23, 2002)
- Saphir, M. N., S. Chaffee, *Adolescents' Contributions to Family Communication Patterns*. Human Communication Research, Jan 2002.
- Saunders, Christopher. *IMing in the Mists of Prehistory*. Instant Messaging Planet. <<http://www.instantmessagingplanet.com/public/article.php/1560461>> (December 20, 2002)
- Small World Research Project*. Dept. of Sociology, Columbia University. March 21, 2003. <http://smallworld.sociology.columbia.edu/>
- Tidwell, L. C., J.B. Walther, *Computer-Mediated Communication Effects on Disclosure, Impressions, and Interpersonal Evaluations: Getting to Know One Another a Bit at a Time*. Human Communication Research, July 2002.
- W3C's Scalable Vector Graphics (SVG)*. World Wide Web Consortium Document Formats, <<http://www.w3.org/Graphics/SVG/Overview.htm#8>>. (November 20, 2002)
- Watts, Duncan J. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company Ltd. London. 2003.
- Woods, Bob. *IM Gaining Popularity in the Enterprise*. Instant Messaging Planet, <http://www.instantmessagingplanet.com/enterprise/article/0,,10816_1469011,00.html> (September 24, 2002)

Woods, Bob. *VeriSign, AOL Teaming on Encrypted IM*. InternetNews.com,

<http://www.internetnews.com/ent-news/article.php/7_1107841> (May 10, 2002)