Mining the Mine

Exploratory Social Network Analysis of the Reality Mining Dataset

Ben Congleton
Satyendra Nainwal

School of Information
University of Michigan, Ann Arbor

Guide: Prof. Lada Adamic

Fall 2007

1 Introduction

The year 2002 marked a turning point in the history of telecommunications. It was in that year that the number of mobile subscribers overtook the number of fixed-line subscribers on a global scale, and mobile became the dominant technology for voice communications. This revolution in communication is more then just a technological advance and has fundamentally changed the way people communicate. This has had profound implications on both how people as individuals perceive communication as well as in the patterns of communication of humans as a society. Indeed, the mobile phone has moved beyond being a mere technological object and become an important part of many people's social lives.

The convenience and ubiquity of the mobile phone are changing the way in which we interact with the information world around us. Mobile phone are no longer simple calling devices, recent innovations have empowered mobile phones to monitor their children, being able to play 'treasure hunt games', access their email, and more recently search the world wide web through a voice interface! With ubiquitous connectivity also comes the ability of cell phones to act as natural sensors. This is increasingly being used to get more accurate data and analysis of group dynamics than was ever possible before. With this context it is both an exciting and an important time to compare mobile social networks with the social networks created through face-to-face and Internet mediums.. In this project we undertake an exploratory social network analysis of the Reality Mining Data from experiments at the MIT media lab.

2 The Data

The reality mining data set was collected over 9 months by monitoring the cellphone usage of 100 participants. 75 participants were MIT Media Lab affiliated professors or students, 25 participants were incoming MIT Sloan freshman. Each participant took a small demographic

survey and was given a Nokia 6600 phone running special logging software to keep track of incoming and outgoing calls, current cell tower id, and any Bluetooth devices detected nearby during the study [1]. Each piece of collected data was stored on the phone during the study and than imported into a MySQL database for analysis. We used an anonymized version of the collected data provided by MIT researchers on the initial Reality Mining Team.

3 Related Work

There has been a significant amount of work on social network analysis of online interactions. Studies have focused on a wide range of issues such as characterizing online interactions [2], effect of the internet on real life interactions [3], understanding real world social networks from online interaction structure [4] and the study of email networks to characterize tie formation [5].

More recently there has been increasingly growing interest in the field of mobile social networks. Google recently bought a mobile networking startup DodgeBall (http://www.dodgeball.com/), Microsoft announced its own mobile social networking tool SLAM ( ), while Helio now features MySpace on its handsets! However, perhaps primarily because of data unavailability, there have been far fewer studies in mobile social networks analysis. The MIT Media Lab reality-mining group and the Context Group Finland at the University of Helsinki have been pioneers in studies in this field and both have made publicly available large datasets from their projects.

4 Motivation and Sources

4.1 Reality Mining Project

In their seminal paper on the Reality Mining dataset the authors describe how they use the cell tower and the Bluetooth proximity data to complement each other and make various inferences about the presence of people. Studying this over a group of people for a long time enabled them to identify structures in the everyday routine of people [1]. Using algorithms and location data over time they were also able to predict, with reasonable accuracy, for low entropy subjects where the subject was likely to be. We however are more interested in studying the community structure in the underlying network of phone calls being made and how that community structure relates to the group distributions in real life (e.g. MIT lab students vs.Sloan students, or first year students Vs. Second year students etc.) We are also interested in studying relationships between the calls being made the location of the people making the call. For example amongst several other things we are interesting in studying how the duration of the calls are co-related with whether people are co-located or not. A more detailed descriptions of our analysis is given later in the paper.

4.2 Community Detection

For the community structure we use the measure of network modularity [6]. Modularity is a property of a network and a specific proposed division of that network into communities. It measures when the division is a good one, in the sense that there are many edges within communities and only
a few between them. Since our original network is significantly large we use the algorithm by Clauset, Newman and Moore to find the modularity and the communities formed [7].

4.3 Social-Structural Determinants of Association

Scott L. Feld [8] in his important work on Social Structural Determinants of Similarity among Associates concludes that the social structuring of activities tends to bring people with similar interests closer to one another. Specifically, he mentions people choosing their friends influenced a great deal by who they are in regular contact with. This was in the age of no computer and mobile phones. After that homophily has been studied in various contexts and various networks including that of a university web social network [4] and a university email network [5]. However, to the best of our knowledge no such analysis has been done for a mobile social network – and so we try to explore this aspect by trying to study the relations between the groups formed by the call network as opposed to the real life groups like members of a research group, or one particular school of people performing a similar kind of activity with reasonably similar goals (e.g second year graduate students). An interesting aspect would have been to study the evolution of the network topology by studying how the network behaves [5], e.g. in terms of closure of the triads, or clustering as the network evolves. However, since our data was only limited to the 100 participants we had no way of knowing the relations amongst the call made by the people who were not in the study but were called by people in the study. This was just one of the many interesting things that we couldn't study well because of that information not being there in the dataset.

4.4 Co-relation between calls and Location

And finally we study correlations between calls the spatial location. The analysis includes studying the relation between the duration of phone calls and the location of the people the call is between. We haven't come across any research that explicitly studies that. Analyzing this can help us glean some insights at a macro level analysis between call duration and the location of the callers, and perhaps lead to some design ideas.

5 Data Analysis

The first stage of analysis focused on using community detection algorithms to understand the community structure of the communication network. In the later stages of our analysis we attempted to understand where study participants called either other, and if there were particular locations were certain study participants were more like to call each other. But, before jumping into this part of the analysis we will detail how we worked with the reality mining data set to build these graphs.

5.1 Mining The data

The data provided by the reality mining team was not in a great format for data analysis. We spent a good deal of time writing a series of scripts to massage the Reality Mining Data into graphs suitable for analysis in Pajek and GUESS.

5.1.1 Tools

5.1.1.1 Custom Scripts

To facilitate data analysis using multiple analysis programs we developed a series of Ruby classes that encapsulated advanced network structures. We than wrote custom importers which let us easily export a view of the network as a series of nodes linked together. For our community analysis nodes were people who were either in the study or called by someone in the study. Each

edge was an aggregation of all incoming, outgoing, and missed calls between the two people it connected.
The ruby classes were also designed to facilitate analysis using Guess & Pajek. Thus, they allowed us to visually manipulate a graph in guess, save it as a GDF, and then convert the GDF to a format suitable for Pajek.
We also wrote a series of short python scripts to help us filter the data in GUESS. For example, we wrote a tool to delete nodes with arbitrary degree, and a program to create one-mode graphs from bipartite graphs.

5.1.1.2 GUESS

Guess is a network analysis tool developed at HP Labs. It has a very powerful console for manipulating graphs, and is full scriptable using Python. We primarily used Guess to build graphs for display, and to gain macro level insights into the network structure.

5.1.1.3 Pajek

Pajek is another network analysis tool. It has a somewhat clunky interface, but has many built in tools for graph analysis, including algorithms to find clustering coefficients, network prestige, and some community finding support. We primarily used Pajek for the algorithmic graph analysis that GUESS lacks.

5.1.1.4 Newman Community Tool

To find communities in our network and characterize the modularity of the network we use the Community finding algorithm for large networks, which is a hierarchical agglomeration algorithm for detecting community structure, by Clauset, Newman and Moore [A. Clauset, M.E.J. Newman and C. Moore, "Finding community structure in very large networks." **Phys. Rev. E 70, 066111 (2004)].**

5.1.2 Data Cleaning

5.1.2.1 Call Analysis

The first step of cleaning the Call Analysis data was to export the call data from the MySQL database into GDF (GUESS Data File) format. This was accomplished using the previously discussed Ruby tools. It is important to note that the RAW reality mining call log data was relatively useless, because the nature of the data collection left many nodes with a single link tying them to a study member. For example, this would occur whenever someone in the study called someone who wasn't in the study. Occasionally two members of the study would share the same outside-of-the-study contact, but this was rare. To mitigate these problems we first removed all the nodes with a degree of 1. This limited our graphs to just people in the study, and people outside the study who were tied to more than one study member. For call analysis we were primarily concerned with the social networks of participants in our study so we divided the graph's nodes into two groups: nodes-in-our study and nodes-out-of-our- study. We projected this bipartite graph into a single mode graph of connections between people in the study.

5.1.2.2 Call Location Analysis

The Reality Mining Dataset lacked an efficient method of connecting individual cellphone calls with the location of the caller and receiver making the call. Thus, to extract this information we had to write scripts to link each participant's phone call records with their location records. Once this was done we used our ruby export classes to build a GDF of people in the study linked by

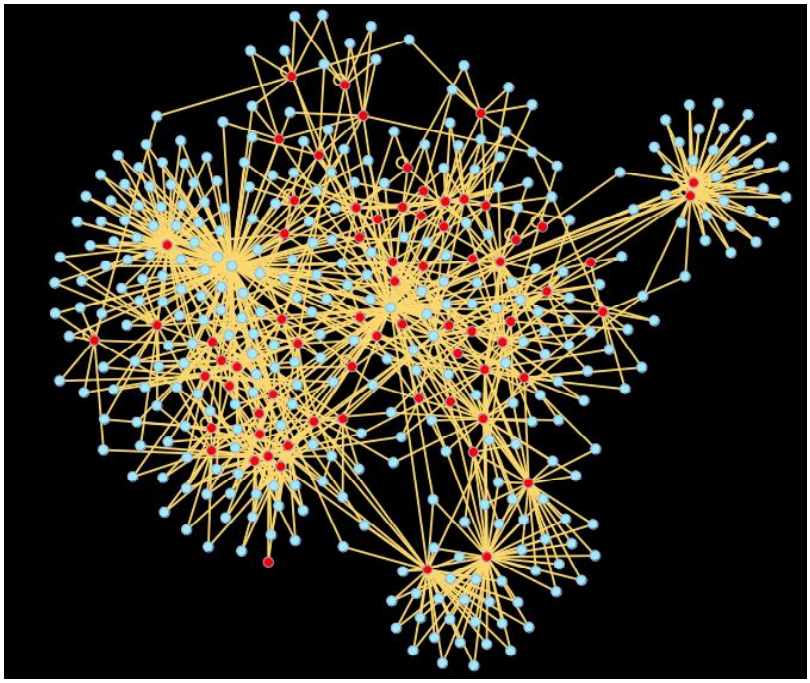
location in which they called each other. For example, if two participants were both in the Media Lab and one called the other, we would draw an edge between them. If one was in the Media Lab, and another was in the Library we would not draw an edge between them. In this way we were able to build networks of participants that made phone calls to each other when they were near each other.
We used a similar method to create a bipartite graph of people making calls to other people at the same location, where nodes were locations and people.
In this section we discuss some of the more interesting graphs that were built during our data analysis.

5.2 Graphs

5.2.1 Call Graphs

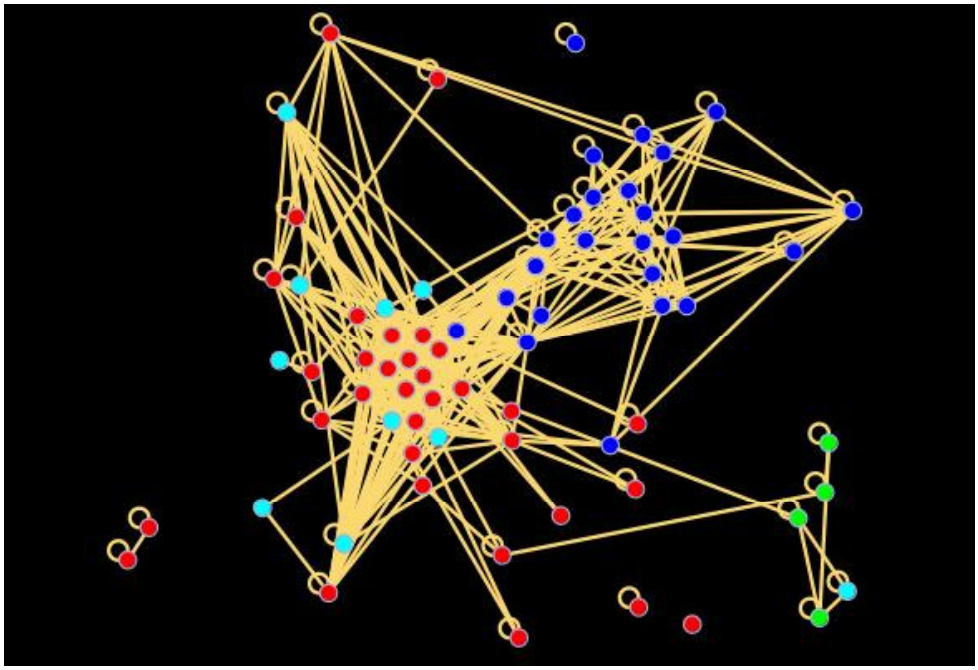
The following graphs were generated by drawing edges between people who either received or made a call during the study period.

5.2.1.1 Bipartite Graph of Study Participants and Non Study Nodes



The red nodes represent study participants, the light blue nodes represent people who either called or received a call from at least two study participants. Edges represent a called or received call from relationship; directional arrows were removed to reduce clutter. The spatial location of the nodes provides insight into groups of people that communicate with each other.

5.2.1.2 One Mode projection of connections between study participants

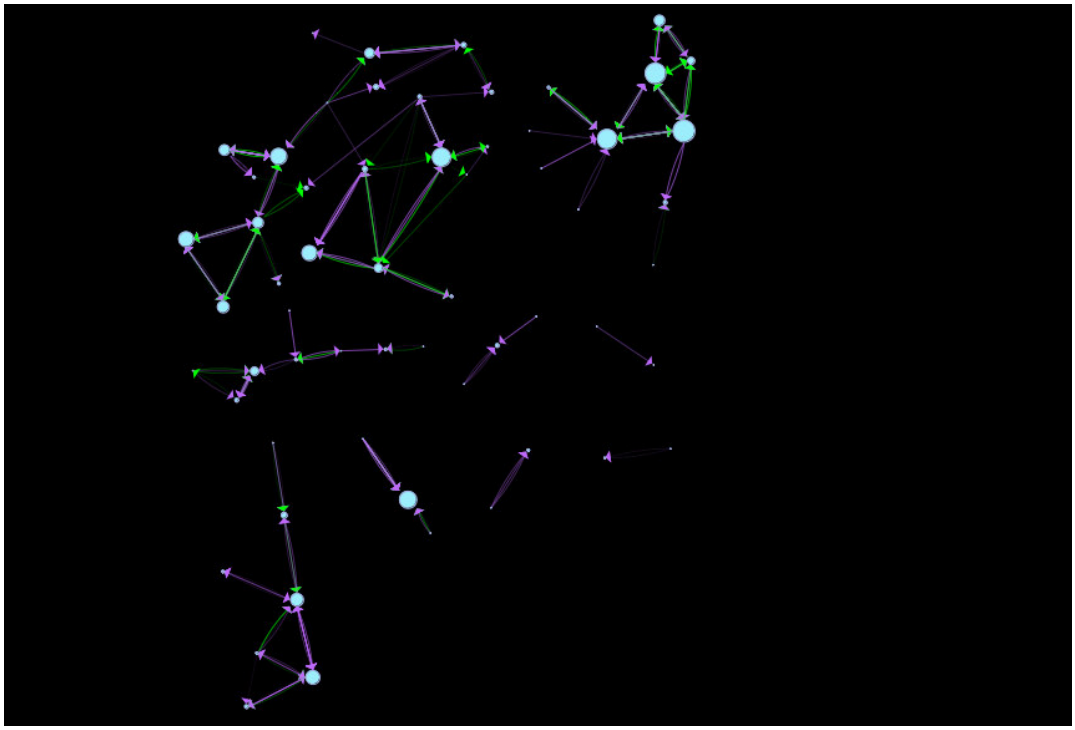


This network is a one-mode projection of the bipartite graph showing all of the connections between study participants. The color of the node represents participant self-identification from the survey results. Green nodes are self-identified as MAS students, blue self-identified as Sloan students, cyan self-identified as Media Lab Students, and the red nodes self identified as graduate students who weren't attached to a specific apartment.

5.2.2 Call Location Graphs

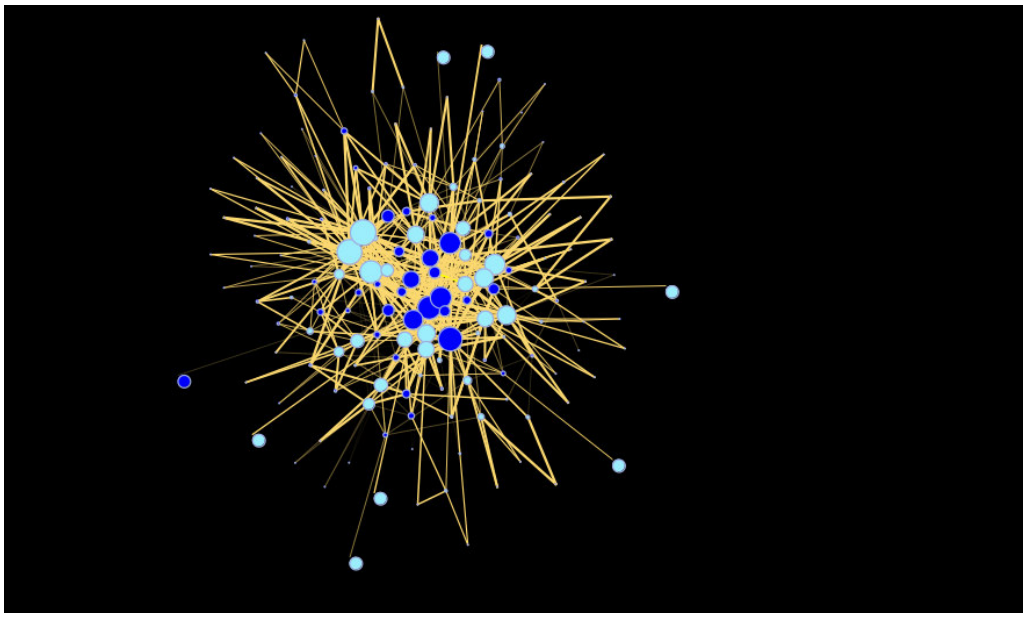
The following graphs were created by examining the localities where one participant called another nearby participant. I.e. we were examining situations like: who calls whom when they are both in the media lab. *[These graphs assume that each cell tower ID is a different location, thus these graphs are only able to capture calls that were made at the same location on the same carrier]*

5.2.2.1 Graph of who calls who when they are co-located



The directed edges in this graph represent outgoing calls occurring between two people at the same location. The thickness of the edge is based on the number of outgoing calls made. Green edges represent calls made in the media lab. Pink edges represent all other calls. The size of the node is relative the degree of the node, thus people who receive calls from many different people appear larger on the graph. It is interesting to observe how some people make many outgoing calls, but receive very few calls, and vice versa.

5.2.2.2 Bipartite Graph of Callers and Locations they call from



Each edge represents a phone call. Dark blue nodes are locations, and lighter blue nodes are people. All nodes are sized based on their degree. Edges are sized based on the total duration of the calls. A phone is represented by an edge drawn between a person and a location, and then another edge drawn between that location and another person. This graph only shows calls that were made and received in the same location. The larger dark blue nodes are locations where lots of different people call each other when they are collocated. (In this example a lot of people call one another when they are in the media lab). Large light blue nodes represent people who call a lot of people when both the caller and the recipient of the call are in the same physical location. The cluster of dark blue nodes near the center of the graph represents cell towers around the MIT Media lab. This tells us that a lot of people in the media lab, call each other when they are both in the media lab.

5.3 Analysis of Communities

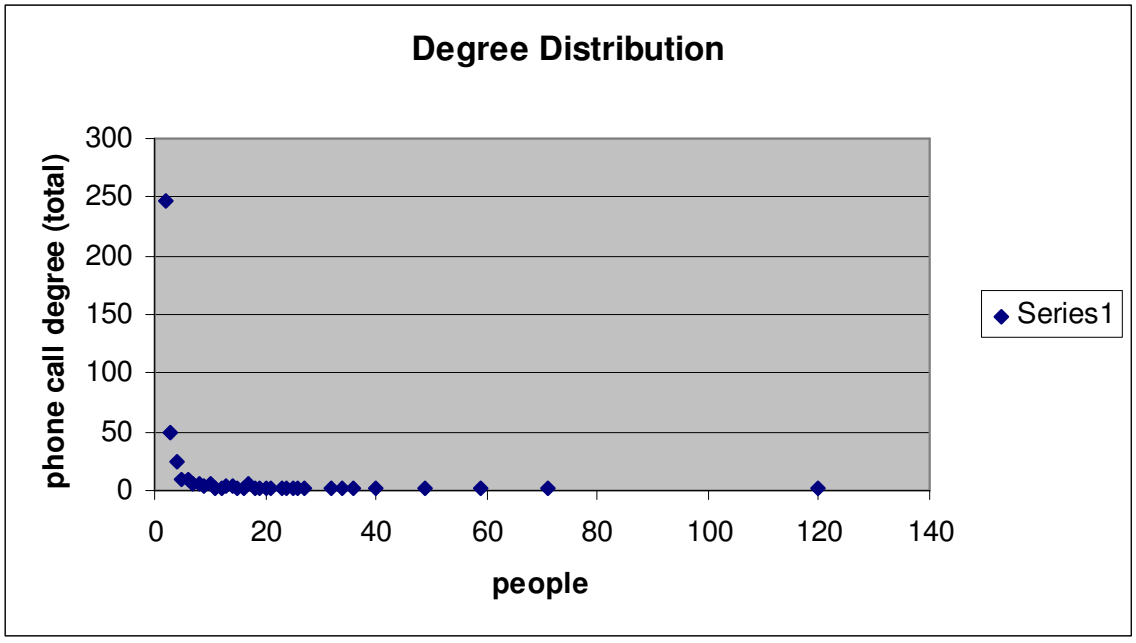
5.3.1 Network Measures

Next we analyzed the data in both Pajek and GUESS and analyzed them, running various network measures on them. A few of the properties that help give an idea about the structure of the graph are given below:

Property	**Reality Mining Data**	**Similar Random Graph**
Clustering Coefficient	**.25**	**.06**
Average Path	**3**	**2.2**
Diameter	**8**	**4**

Thus it is a small world network with a high clustering coefficient and small average path.

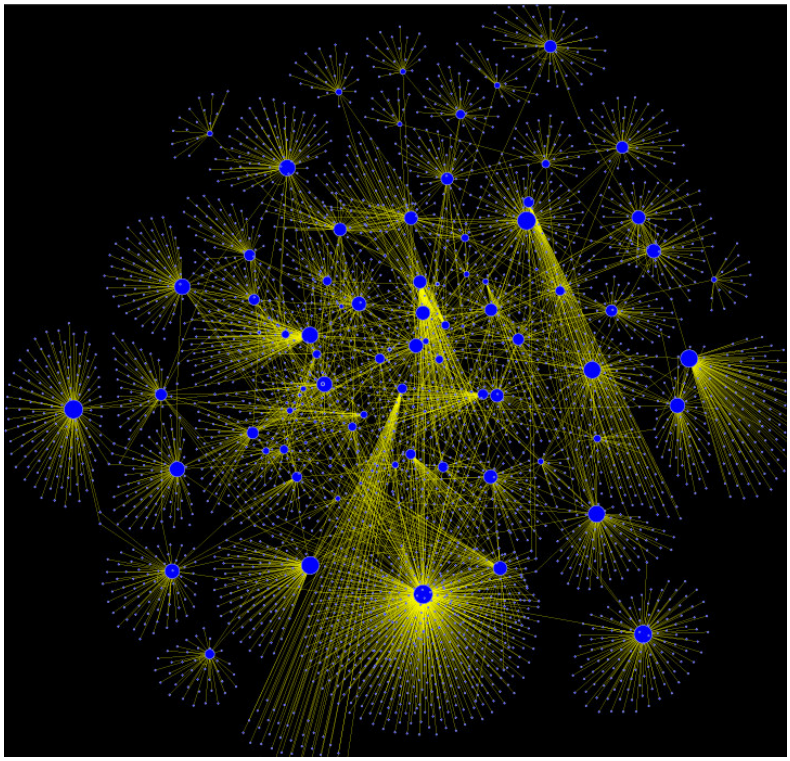
The degree distribution of the graph follows a power law distribution.



5.3.2 Communities in the Complete Dataset

We started with running the Clauset-Newman modularity algorithm on our dataset and got an exceptionally ***high modularity of 0.85 , with 34 groups***!

When we looked at the network structure of the graph, however, it was clear why we had got such a high value of modularity for the graph (shown below)



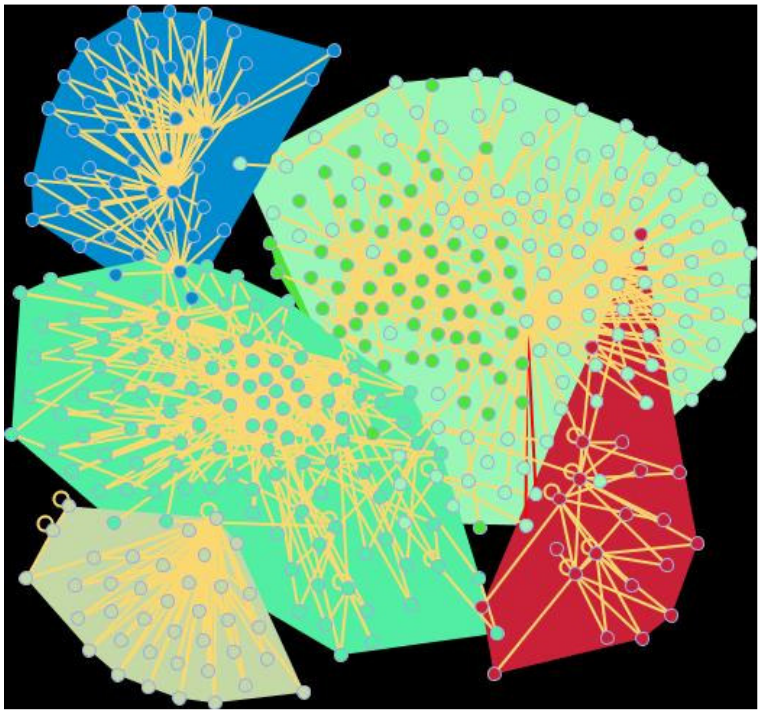
From the graph it was clear that the high community structure was due to the participants calling a large number of people other than the 100 participants. A good number of the communities were formed around the participants calling people outside. Since we did not have any data on who the people being called were calling there was no way we could get any meaningful results about the clustering coefficient or triad closures about the graph. To make better sense of the calling pattern of participants in the study depending on what we knew about them from the calling logs we now proceeded to trim down the leaf and low degree nodes to focus only on the participants and the calls made between the hundred people in the study.

5.3.3 Communities in the Network of Participants

On running the Newman Modularity algorithm on the smaller graph of participants only we got a ***modularity of 0.34*** *and the following group distribution.*

Number of groups	4
Minimum Size	4
Mean Size	17
Maximum Size	29

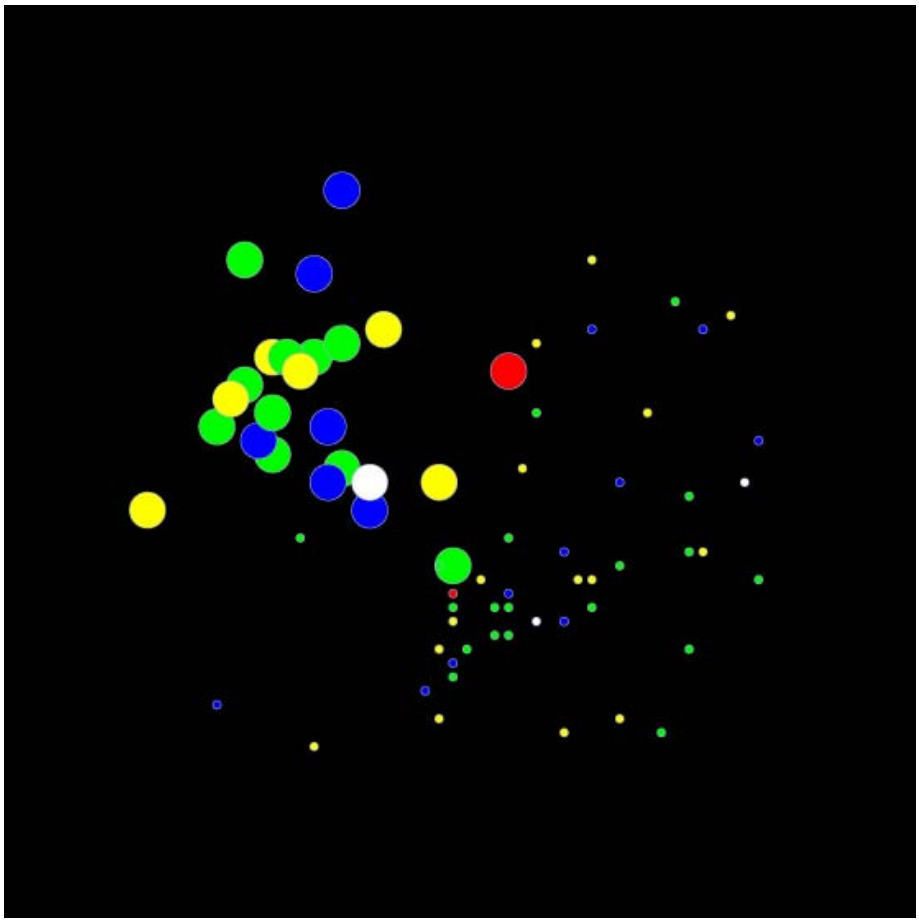
We wanted to study how the frequency of the people calling the other participants in the study compared with the real life social structuring of activity around them. To do this we divide the participants into groups in two ways. In the first grouping we group people who are likely to be physically closer to one another and pursuing similar professional goals together. In this case we get groups like the Sloan Students and the MIT media lab students. For the second grouping we divide the participants into groups based on the groups returned by the Newman Modularity algorithm, which basically groups them together such the number of calls made between the participants in a group is significantly higher than the number of calls made between two different groups and is greater than that expected by random. We also ran the betweenness clustering algorithm on our smaller dataset – participants only- and got results similar to that by the modularity algorithm.



5.4 Comparing Real World Groups with Call Network Communities

To get a visual idea of the way the two different methods of grouping were related we followed the following procedure. We first colored the different groups got from the Modularity algorithm differently. Thus, since there were 4 different groups from the algorithm we had nodes with 4 different colors, each color representing one group. Next we studied the pattern of calls being made between the groups.

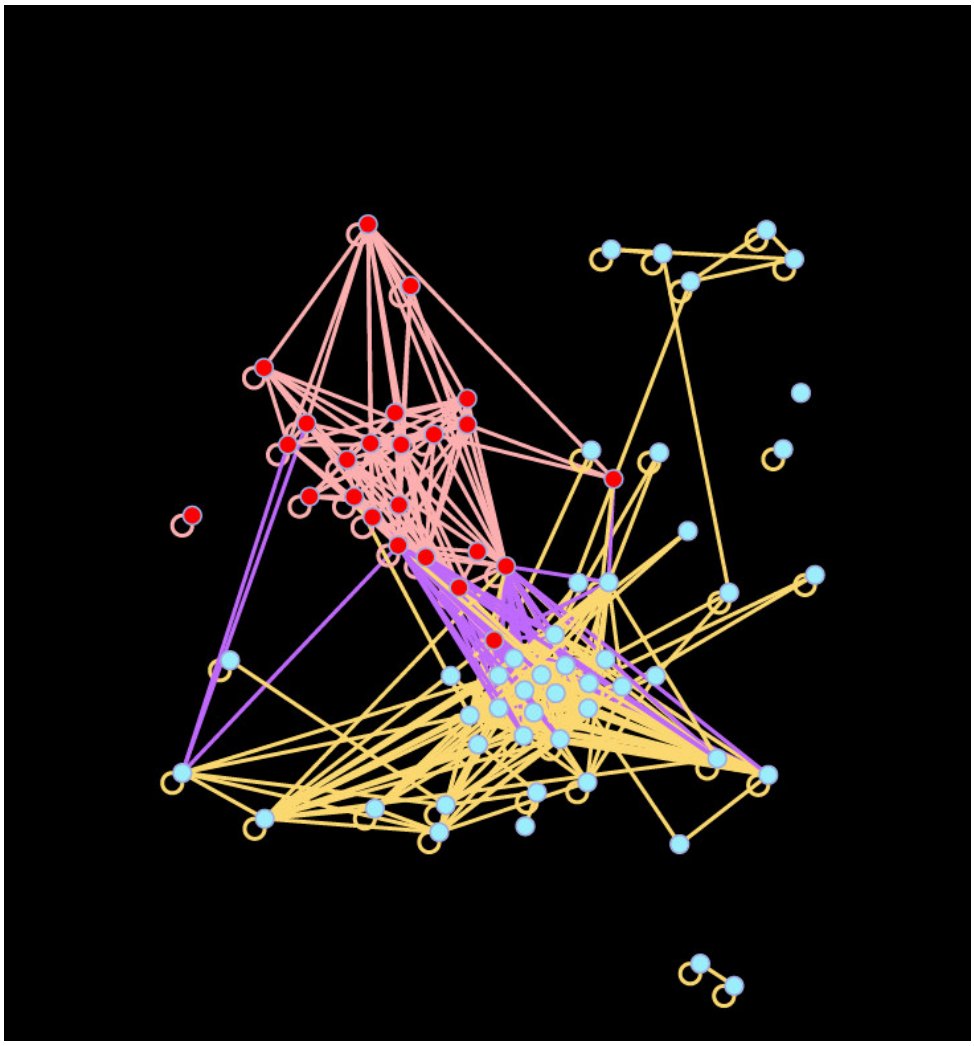
In the image below, all the Sloan nodes are made bigger to differentiate them from the MIT media lab participants. Clearly form the graph, the Sloan students belong to three different modularity groups!



The Sloan students are the larger nodes
The colors denote different groups from modularity algorithm

We can see that amongst the study participants the groups formed by the frequency of calls are not the same as the organizational groups the students are in – e.g. the large nodes are all Sloan school students but as is clear from the different colors these participants belong to three different groups in the modularity analysis.

Thus clearly there are other factors apart from spatial proximity, by virtue of them all being in the Sloan Business School, and considerable amount of shared set of activities that causes people to associate to a considerable degree with other not necessarily in their foci of activity as compared to those in their foci of activity. Feld mentions that "whatever the basis of their initial association with a focus, it may be difficult, costly and time consuming to dissociate from the focus and/or become associated with others". Thus, people tend to choose friends from amongst those with whom "they have regular contact in one or another focused activity". However, ***clearly in the dataset above people have chosen to associate with and continued their association over an academic year with other who are not in their foci of activity. One reason for this may be the increased ease due to instant communication aids such as messaging and mobile phones with which people can now stay in touch with one another – which doesn't make staying in touch with people not in your foci costly anymore.***
The following graph nicely illustrates the high amount of inter-group calls between the Media Lab and Sloan students.



In red are the Sloan participants and in Blue the Media Lab people. The purple edges denote the communication that took place between the two groups.

6 Conclusion

From our exploratory network analysis there were some interesting insights we gained into the dataset and some interesting patterns which have great potential for a more rigorous and analytical exploration.

- Feld mentions "it may be difficult, costly and time consuming to dissociate from the focus and/or become associated with others", however with the difficulty time and cost of communication with people farther away becoming continually smaller we may see more of group formation even outside the Foci. Thus Foci of Activity or organizational structure may not be the only indicators of community formation in networks like this.

- The ***total and average duration*** of calls made when people were not at the same location was higher, however there were far more calls of median duration when people were co-

located. This may possibly be very interesting information about call behavior of people and something worth pursuing in greater detail and maybe compare with another similar dataset.

- We found an interesting pattern in call reciprocity in the call network. In instances of the call network where there was a triad, i.e. when people called two others when they were in the same location as the other person, the mutual exchange of calls or call reciprocity was higher than when two people mostly called each other when they were in the same location. In the latter case the call exchange was highly asymmetric with one person calling the other most of the time.

In the near future we expect to conduct a more rigorous analysis of findings two and three and perhaps compare these results with another mobile social network to see whether these results are a phenomenon particular to this network or perhaps a more general property of mobile social networks.

7 Acknowledgements

We would like to thank Prof. Lada Adamic for her constant help and guidance and Nathan Eagle for so readily providing the Reality Mining Dataset.

8 References

[1] N. Eagle and A. Pentland (2006), "Reality Mining: Sensing Complex Social Systems", *Personal and Ubiquitous Computing*, Vol 10, #4, 2006.
[2] P. Curtis, 1992. "Mudding: Social phenomena in text-based virtual realities," In: *Proceedings of the 1992 Conference on the Directions and Implications of Advanced Computing,* Berkeley, Calif. (May).
[3] S. Wasserman and K. Faust, 1994. *Social network analysis.* Cambridge: Cambridge University Press, pp. 188-191.
[4] L. A. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the web. First Monday, 8(6), June 2003.
[5] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. Science, 311(5757):88-90, January 2006.

[6] Modularity and community structure in Networks, M. E. J. Newman, PNAS | June 6, 2006 | vol. 103 | no. 23 | 8577-8582

[7] A. Clauset, M.E.J. Newman and C. Moore, "Finding community structure in very large networks." Phys. Rev. E **70**, 066111 (2004

[8] S. Feld. Social structural determinants of similarity among associates. American Sociological Review, 47, 1982.

[9] Watts & Strogatz, 'Collective Dynamics of Small World Networks', Nature, 1998

[10] L. A. Adamic and E. Adar. Friends and neighbors on the web. Social Networks, 25(3):211{230, 2003.
[11] N. Eagle, A. Pentland, and D. Lazer (2007), "Inferring Social Network Structure using Mobile Phone Data", *PNAS*, (in submission).