Discovery of User Groups within Mobile Data

Syed Agha Muhammad
Technical University Darmstadt
muhammad@ess.tu-darmstadt.de

Kristof Van Laerhoven
Darmstadt, Germany
kristof@ess.tu-darmstadt.de

ABSTRACT

Groups of individuals in terms of social structure and behavior can be revealed in location and social interaction data from their smart phones. These data can be cross-analyzed to find common contacts and landmarks across users, which are called and frequented at approximately the same time. In this paper, we present a graph-based approach to model and identify significant groups of users by analyzing their mobile phone data. We propose the use of 5 modalities to create these models: shared contacts, shared IDs on the call lists, and common Bluetooth and WLAN MAC addresses seen by the users, and GPS clusters in close proximity in time and space from each other. An evaluation on data from 37 users from the same city shows that for the five models individually, similar groups emerge.

1. INTRODUCTION

Over the last few decades, there has been an enormous amount of attention given to social networks analysis (SNA) as key determinant in modern sociology [8]. The basic unit of SNA is an entity that consists of a collection of individuals and the relationships amongst them. Traditionally, the data about the social interaction are gathered using labor-intensive methodologies, which are time consuming and constrained by limited people monitoring. But more recently, the induction of mobile phones in the study of social data collection are gaining grounds. The penetration of the mobile phones has surged in the last 10 years. Mobile technologies are now equipped with large number of built-in sensors such as accelerometer, Bluetooth, GPS, gyroscope etc., that can be used to collect over the longer periods of time. The longitudinal data collected from the wide range of sensors can be used for human behavior modeling.

Apart from individual behavior modeling, a difficult challenge is to discover the complex social structures for collective behavior modeling. In this study, we argue that groups of individuals in terms of social structure and behavior can be revealed in location and contact data from their smart phones. These data can be cross-analyzed to find common contacts and landmarks across users, which are called and frequented at approximately the same time. A group is defined as "A number of individuals assembled together or having some unifying relationship". Detection of groups and their location will make the detection of events possible, especially when such events are semantically meaningful in terms of overall actions of multiple persons considered jointly but not individually.

In this paper, we are interested in analyzing the different data modalities to discover meaningful groups from them. A key challenge lies in the fact that data provided are unlabeled and do not provide any ground truth about the users knowing each other. We use the social interaction and location data from the rich data set provided to us. From the social interaction data, we analyze only Bluetooth scanning results, call logs and contact data of the users. Similarly from the location data category, we analyze WLAN and GPS data. We use certain attributes from each of the modalities. The selected attributes are: call list, shared contacts, Bluetooth MAC address, WLAN MAC address and GPS cluster. To discover the meaningful groups, we have used the graph based approach for all 5 modalities. Similarly, depending upon the information shared amongst the users, we have developed a technique to discover the meaningful groups.

The remaining of the paper is structured as follows: Section II reviews the related work on social data collection using mobile phones, discovering meaningful places using GPS data and forming groups using Bluetooth. Sections III discuss the methodology for modeling the modalities. It presents an introduction to the Nokia mobile data challenge (MDC), and discusses the methodologies applied on different modalities for discovering meaningful groups. Sections IV discuss the experimental results of the call log, contacts, Bluetooth, GPS and WLAN data. We used the node based approach to detect the meaningful groups. Section V discuss the main conclusions of this paper.

2. RELATED WORK

Mobile phones are nowadays extensively used for the collection of contextual data of the users to construct a social network. Reality mining project [5] has conducted an extensive research on the use of mobile phone to provide insight into the dynamics of both individual and group behavior. However, their research focusses mostly on human behavior modeling [3], social network analysis[9], and human mobility analysis [6].

Apart from, a rich content of research have focused on using single modalities GPS [1], Bluetooth [2], or a combination of contacts and Bluetooth data [4] to discover the meaningful places and groups across the users. In [1], authors have modeled an application to detect the meaningful places visited by multiple users. Different users can query the data of the other users to find if they can meet each other at a specific place. Bluetooth data can be used to analyze prox-

imity interaction to discover the temporally grounded social context, normally like being at home or at a meeting. In [1], the authors have used a probabilistic model for real-life social context discovery. In [4], the authors have used the combination of Bluetooth and call log data to construct a social network. The data was collected from the smart phones and self-reporting technique.

3. DATA SET AND METHODOLOGY

This section presents the data modalities, and methodology used for discovering the groups.

3.1 Data Set

The data is taken from Lausanne Data Collection Campaign (LDCC) [7]. We analyzed the data from 37 different users. For our research, we used the call log, contacts, Bluetooth, GPS, and WLAN modalities [1]. Table 1 summarize the attributes used for our analysis.

Groups	Attributes
Call log	userid, tz, call-time, direction, number (anonzmiyed number)
Contacts	userid, first-name, last-name
Bluetooth	userid, time, tz, mac-address
GPS	userid, time, tz, longitude, latitude
WLAN	userid, time, tz, mac-address

Table 1: Data types used for analysis.

3.2 Methodology

The following subsections describe the methodology that we employ to model different modalities. We followed the same procedure of using the MAC address as a parameter to get the results from the Bluetooth and WLAN data set. Instead of making a separate subsection for the WLAN, we have combined it with Bluetooth.

3.2.1 Call-log

Call-log data provide the list of the incoming and outgoing calls dialed by the user. For our analysis, we incorporate the following weighting factors: 1) Same numbers dialed by multiple users. 2) Number of times the same anonymized number is found in the data for multiple users. 3) Direction of the call.

The above mentioned three factors were combined to detect the meaningful groups. In everyday life, people mostly call or message those people with whom they have acquaintanceship. Following this assumption, we checked the common numbers and the number of times they appear in the call logs.

3.2.2 Contacts

Contact data provide the contact lists of the user. We select the number of common contacts shared by the users as a weighting factor for modeling. For some entries in the data set, first-name or the last-name was missing. We have not considered data for results.

[1]For detailed discussion about the attributes, check the Referred document

3.2.3 Bluetooth

The procedure for detecting groups was similar for both the modalities. Bluetooth data provides the number of devices seen by the user. Due to the short-range nature of Bluetooth communication, it can predict with a high probability when two users are in close vicinity of each other. It is also a very crucial modality to capture the social interaction between the users. Our purpose was to track the common Bluetooth devices, which sense the approximation of face to face communication. We analyzed the data using following two steps: 1) Initially, the common Bluetooth devices seen by the users were found. 2) Cross analyzed the timestamps occurrences of the common Bluetooth names, to verify that these users were seeing common Bluetooth devices at the same time.

Due to energy constraints and other issues, data was missing at points. We discovered in the data set that sometimes two users see common devices with the difference of a certain time period, which is within 30 minutes, and then the Bluetooth data for one of the users gets missing only to reappear after some time. In order to decide what value of time difference to choose to get the maximum number of common Bluetooth devices, we performed series of experiments by changing the time difference within the range of one hour. Fig 1 shows the number of common Bluetooth devices seen by the users with the time difference. In Figure 1, x-axis and y-axis represent the time difference and number of common Bluetooth devices seen by the users respectively. The red lines represent the medium that is increasing with time.We can clearly see in the figure that when the time difference increases, the number of times common devices see each other also increases. This indicates a linear relationship between the two. In the end, we decided 30 minutes as an amount of time to track the common Bluetooth devices. Same procedure was repeated for GPS and WLAN data.

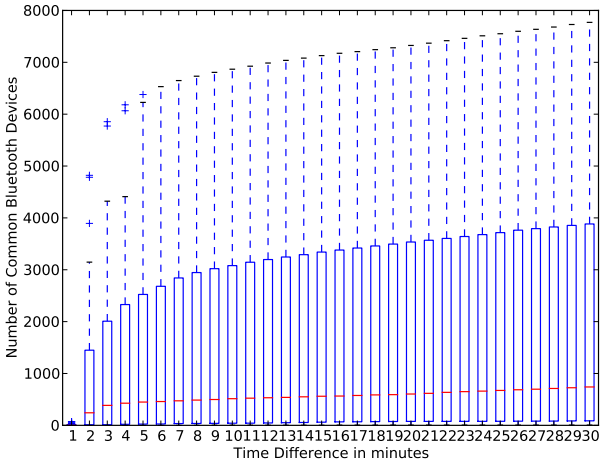


Figure 1: Number of common Bluetooth devices found for varying the time difference.

3.2.4 GPS

To analyze the GPS data, we discarded the noise entries from the data. Our goal was to group users that are in close proximity of each other. We performed the following steps: 1) Compared the timestamps of the users. 2) Selected the latitude and longitude data of the users with minimum time

difference.

After selecting the points, we used the following haversine formula to calculate the distance between two points:

$$Diff\_lat = latitude2 - latitude1 \quad (1)$$

$$Diff\_longit = longitude2 - longitude1 \quad (2)$$

$$a = (\sin(Diff\_lat/2))^2 + (\sin(Dif\_longit/2))^2 * ((cos(latitude1) * \cos(latitude2))) \quad (3)$$

$$c = 2 * \arctan(\sqrt{(a)}, \sqrt{(1-a)}) \quad (4)$$

$$Distance = R * c \quad (5)$$

Where lat, longit denote the latitude and longitude and R denotes the earth's radius. The values of latitude and longitude were converted to radians before using them. The radius of the earth is 6371 km. Figure 2 shows the results for the GPS data for 37 users. The x-axis represents the distance in meters and y-axis represents the number of distance points for all the users. The chart shows that as the distances between the users increase, the number of common points amongst the users also increases.

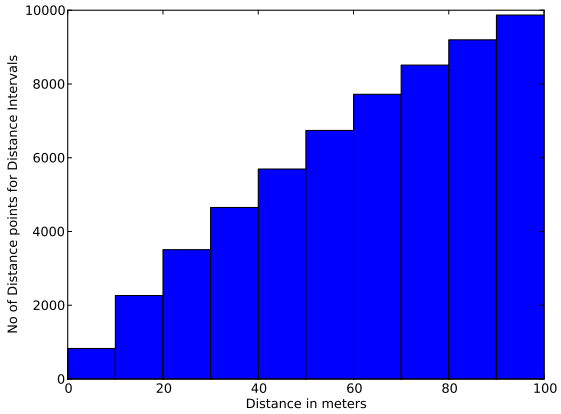


Figure 2: Data Distribution for the GPS data.

4. RESULTS

Figure 3 shows the data distribution across all the modalities. The purpose of the histogram was to find the visual impression of the data distribution in the modalities.

To extract groups across the modalities, we used certain percentage of the probability distribution of the data. The selected percentages started from 10% till 40%. Figure 4, 5, 6, 7, 8 shows the node diagram for the difference percentage of data distribution. The similar meaningful groups discovered within different modalities are represented by the nodes with the same color. In figure 5, and 6 nodes 120 and 2 are colored orange, because these nodes are shared with multiple groups.

In order to discover meaningful groups across different modalities, a clustering technique is developed. This clustering

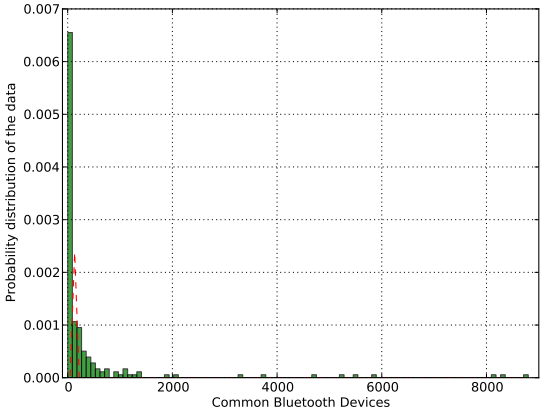


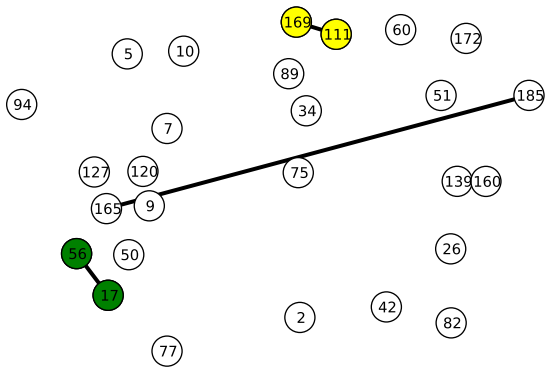
Figure 3: Distribution of the data.

technique utilizes a top-down approach to form groups. For each node we found all its adjacent nodes, i.e., the nodes which are directly connected with it. In Figure 6a, we selected node 9 as our first node and found all the nodes connected to it (50,120, 75, 42, and 2). In the next step, we find the nodes which are connected to the adjacent nodes, i.e., second level neighbors of a node. In figure 6a, node 50 is connected is 126, similarly node 75 with node 2 or vice versa, also 42 is connected with 75. This procedure is repeated until there are no nodes left. As we can see in figure 6a that 3 groups (120, 9, 75 and 9,75,2, and 9, 75, 42) has emerged. The groups were merely formed amongst the nodes which had a direct connection in-between. This procedure was repeated for all 37 nodes and groups across them were discovered. At the last stage, we merged the groups that shared maximum common nodes. For the case discussed above, we can see that these groups shared at least 2 common nodes in-between. In the last stage, we merged the groups having maximum number of common nodes to form a bigger group.

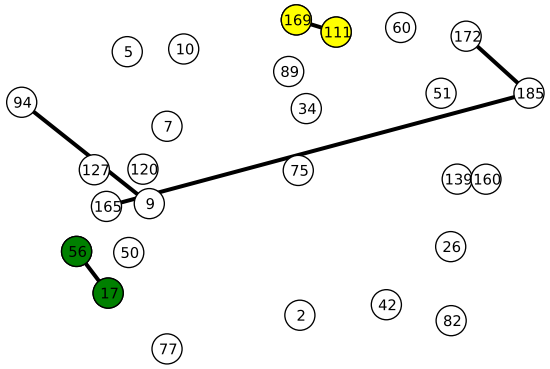
From the figures, we observed that by increasing the data percentage the number of emerging groups also increases. For contact modality, we discovered a single group. Similarly, for Bluetooth, GPS and WLAN modality 3 groups have emerged. Since the provided data is rich in contents, and there always exists the possibility that the formed groups does exists and their data has just matched accidentally at some points. The modeling of different modalities encouraged us to minimize those likelihoods. Additionally, we checked for the number of matched events in the data for different days. There we observed that the users were seeing common devices not just once but on multiple days.

We found the groups that were common or atleast had some common users within different modelaties. Tables 2 summarizes the similar groups for 5 modalities for 40% data distribution. An interesting relationship was found for the strong Bluetooth and WLAN edges. The results have many similarities, which possibly hint that those users were in close proximity of each other. Normally Bluetooth devices have line of sight communication and on top of it seeing common WLAN devices possibly hint these users are interacting with eachother.

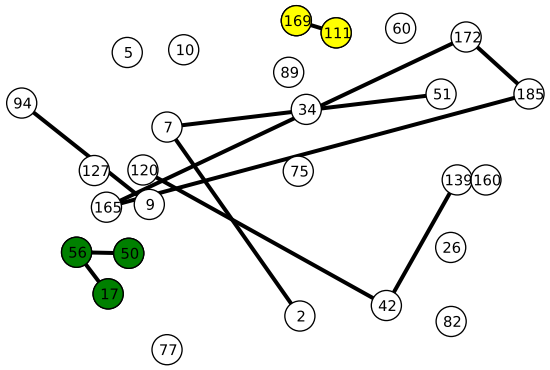
5. CONCLUSIONS AND OUTLOOK



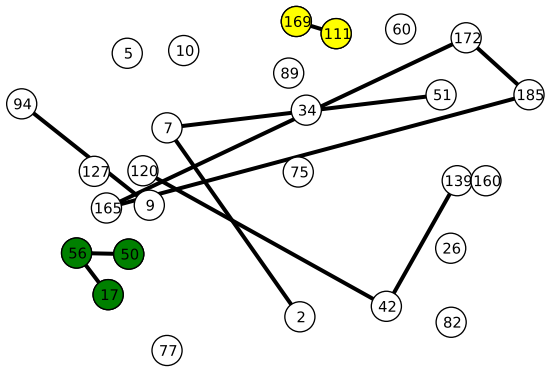
(a) 10% of the data distribution (2 small groups)



(b) 20% of the data distribution (node 172 added)

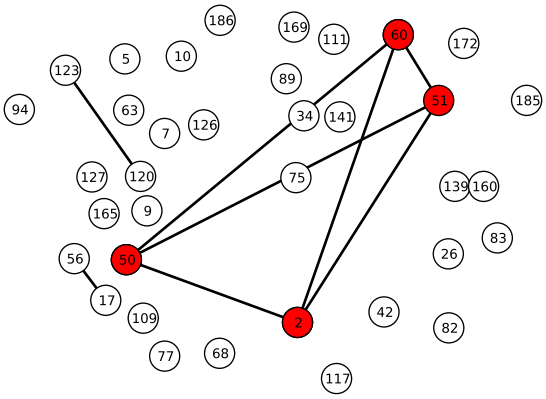


(c) 30% of the data distribution (groups: 161,111 and 50,56,17

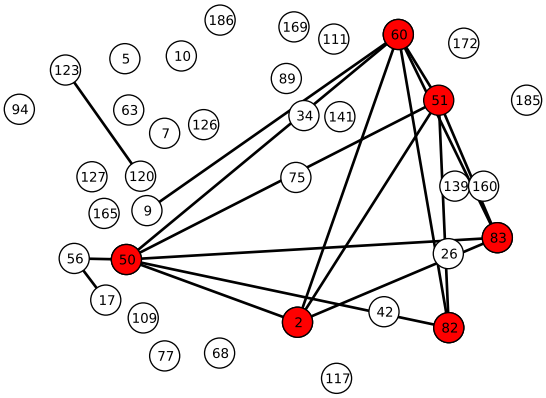


(d) 40% of the data distribution (similar as 30%)

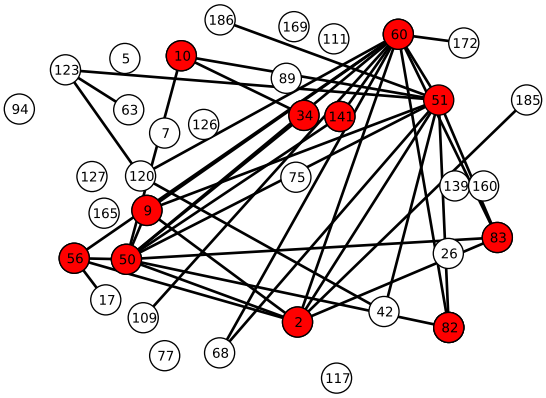
Figure 4: Call-log diagram for different percentage of data distribution.



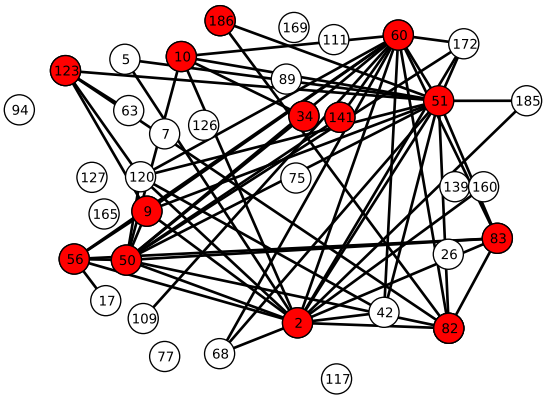
(a) 10% of the data distribution (single group with 4 members)



(b) 20% of the data distribution (node 82 and 83 added)

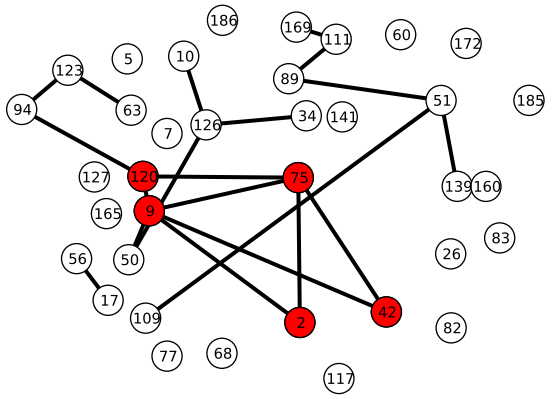


(c) 30% of the data distribution (node, 9,10,34,56,141 added)

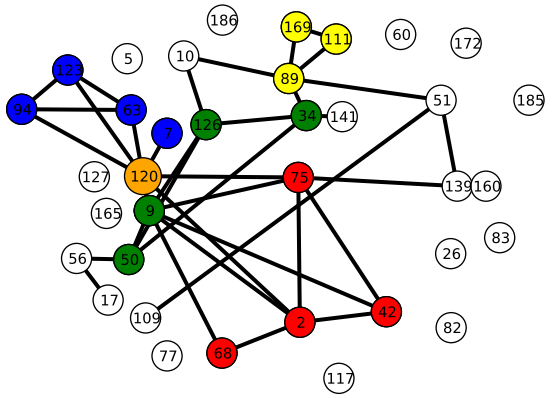


(d) 40% of the data distribution(node 128,186 added)

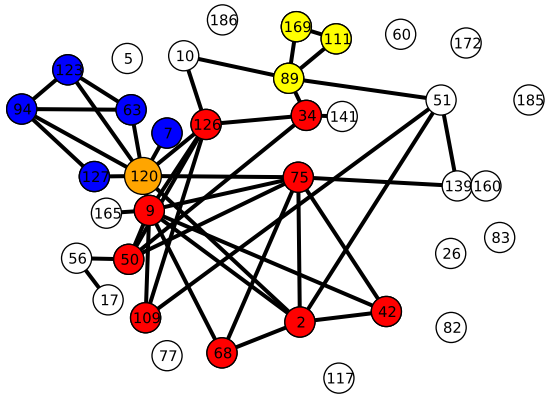
Figure 5: Contact Node Diagram for different percentage of data distribution.



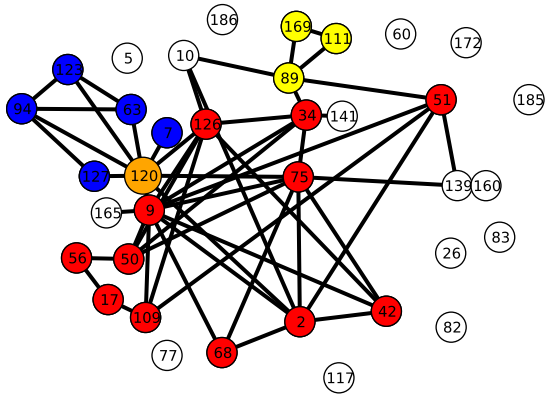
(a) 10% of the data distribution (single group with 5 members)



(b) 20% of the data distribution (4 different groups)

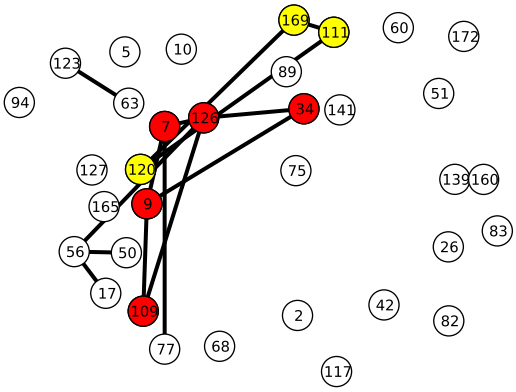


(c) 30% of the data distribution (two groups have merged that results in a 3 groups)

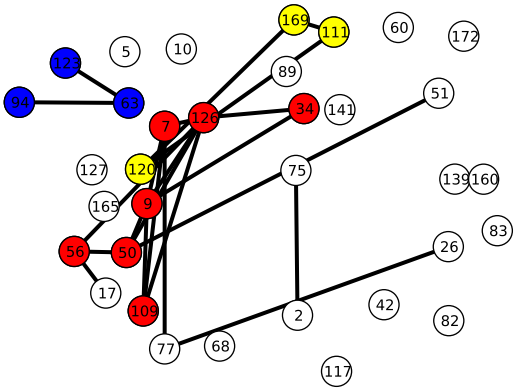


(d) 40% of the data distribution(3 groups emerged)

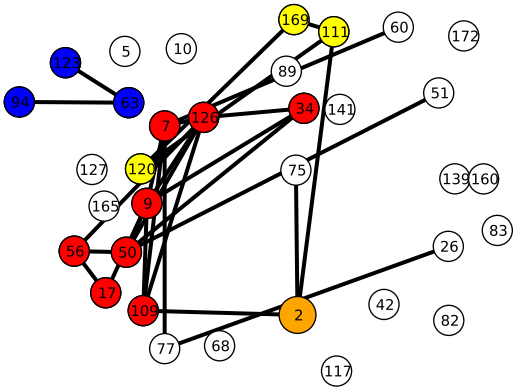
Figure 6: Bluetooth Node Diagram for different percentage of data distribution.



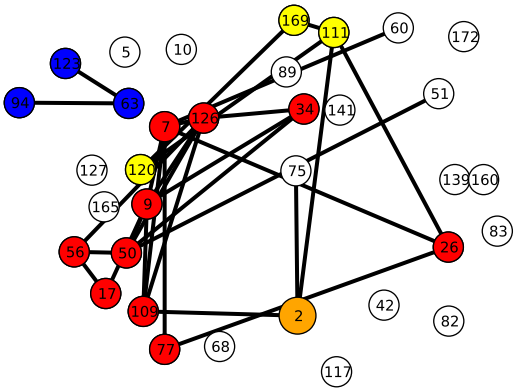
(a) 10% of the data distribution (2 groups)



(b) 20% of the data distribution (existing groups have widened and a third group has emerged)

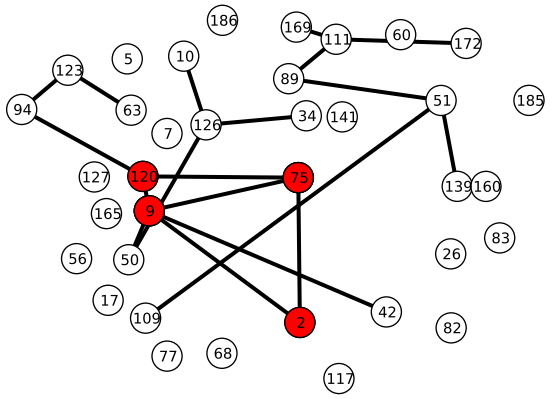


(c) 30% of the data distribution (node 17 and 2 added with the existing group)

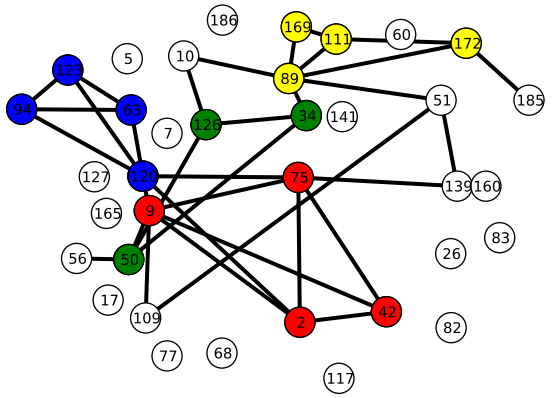


(d) 40% of the data distribution(node 77 and 26 added)

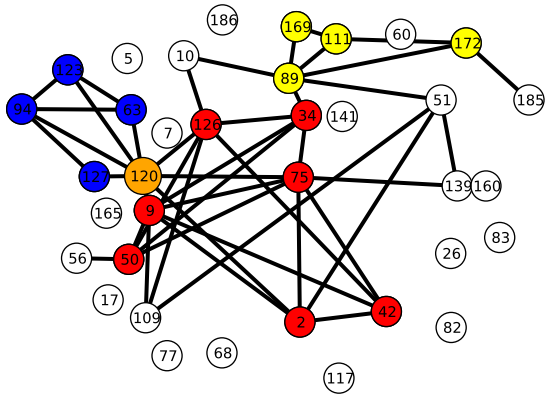
Figure 7: GPS Node Diagram for different percentage of data distribution.



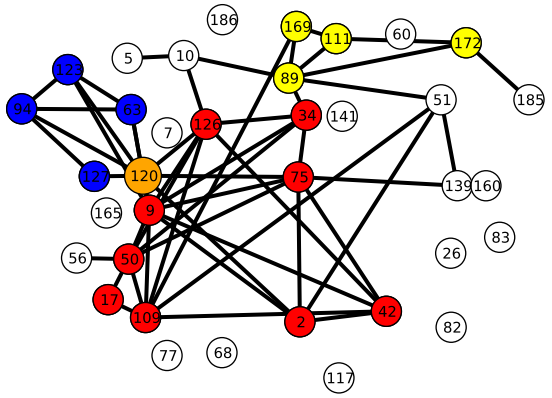
(a) 10% of the data distribution(single group)



(b) 20% of the data distribution (4 groups have emerged)



(c) 30% of the data distribution



(d) 40% of the data distribution (3 big groups have emerged)

Figure 8: WLAN Node Diagram for different percentage of data.

Groups	G1	G2	G3
Call log	-	111,169	17,50,56
Contacts	50,56,9,10 2,82,83,51, 123,34,141, 60, 186	-	
Bluetooth	63,94,120, 123,127	89,111, 169	17,50,56,51, 109,9,68,120, 2,126,34,75,9
GPS	63,94, 123	111,120, 169,2	50,56,109,77,9 17,7,126,34,26
WLAN	63,94,120, 123,127	89,111, 169,172	17,50,109,2,42 34,126,75,9

Table 2: Similar groups in 5 modalities.

In this paper, we propose detection of groups of users within mobile data by analyzing 5 individual modalities. Although we do not have a ground-truth on which users actually formed social groups, there is a large amount of overlap between the modalities which can be seen as promising. The experimental results have shown that there are some common groups across all the modalities. We discovered 3 user groups with similar behavior from the considered modalities.

In the future, we would like to improve our group discovery technique. We would also like to find inter-modality groups.

6. REFERENCES

[1] D. Ashbrook and T. Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput.*, 7(5):275–286, Oct. 2003.

[2] T.-M.-T. Do and D. Gatica-Perez. Contextual grouping: discovering real-life interaction types from longitudinal bluetooth data. In *12th International Conference on Mobile Data Management*, june 2011.

[3] N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. Technical report, IN PROC. OF UBICOMP'06, 2006.

[4] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15274–15278, 2009.

[5] N. Eagle and A. (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, Mar. 2006.

[6] P. Hui. Human mobility models and opportunistic communication system design. 2008.

[7] N. Kiukkonen, B. J., O. Dousse, D. Gatica-Perez, and L. J. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ACM Int. Conf. on Pervasive Services (ICPS,',',), Berlin.*, 7.

[8] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Pres, 1994.

[9] E. Yoneki, P. Hui, and J. Crowcroft. Visualizing community detection in opportunistic networks. *Proceedings of the second ACM workshop on Challenged networks*, pages 93–96, 2007.