Discovering Local Social Groups using Mobility Data

Vishnu Jayadevan
B. Tech
CSE
SRM University
Chennai, India

Kinshuk Bhardwaj
B. Tech
CSE
SRM University
Chennai, India

Anshu Kumar
B. Tech
CSE
SRM University
Chennai, India

Prateek Khandelwal
B. Tech
CSE
SRM University
Chennai, India

ABSTRACT
The ever increasing popularity of location aware mobile devices, has facilitated the collection of large amounts of human mobility data. This data usually contains where a person was and when, which can give us an insight into the complex social behavior of people. In this paper we explore the use of co-location as a means to estimate the strength of relationship between people. First we modify existing co-location algorithms to take the significance of the location of co-occurrence into account in predicting relationship strength. Next, we propose a system to identify social groups that exists in geographic areas called Local Social Groups by using co-location information. Finally, we run extensive experiments on our sample data to test the efficiency of our approach.

Keywords
Social group discovery, mobility data, location aware mobile devices

1. INTRODUCTION
The vast quantities of spatiotemporal data available at our disposal provide a sneak peek into the complex social behavior of people. Mobility data is available from a vast spectrum of sources like location based social networks like Foursquare [1], location tagged content on social networking sites like Facebook [2] and Twitter [3]. Foursquare and most location based social networks rely on user checking in to a place. These check-ins are quite important because people check-in to places of particular interest to them. Other normal social networking services like Facebook also contain location data which are tagged with photos or status update similar to check-ins in location based social networks. Another source of location data is through photo sharing sites like Flickr [5] that contain images with location data embedded in image metadata and user description. Many studies have explored the efficient extraction of location data from Flickr [4].

Current online social networks mostly use interaction data generated online by the user as an indicator of the strength of relationship between users. However this system often is error-prone and inefficient. Predicting relationships using mobility data would add another dimension to the current social network analysis engines run by most social networks like Facebook's Graph Search engine [6].

Current research on using co-location as a relationship metric works on the assumption that, your relationships are often the result of a shared activity [7]. This activity can happen over the online or offline environments. In this paper, we are focused on the offline environment where shared activities are focused around a geographic location. That is, more the number of times two people are found together at different places at different times, the more it suggests that there is a significant relationship between the two. This however does not work well in a practical application as people working in the same offices or studying in the same college would show a very strong relationship even though they may not know each other. This requires that the significance of the location where the co-occurrence occurred must also be considered in determining the strength of a relationship. In this paper we use filter the data based on the significance of the location of co-occurrence.

In this paper we also strive to identify Local Social Groups (LSG) which are a tightly knit group of people with strong relationships among themselves. Identifying such groups have a lot of application in different areas. For example social groups can be used in targeted advertisement systems where instead of focusing merely on the individual's likes and dislikes advertisers can now target the group as a whole as well. Identifying LSGs can also be of use in setting up custom privacy filters that in online social networking services selectively. Contacts on online social networking services can be categorized into distinct social groups similar to how Google+ circles work. Privacy protection would be vastly improved if such custom privacy filters can be setup based on how strong your relationship with a person is.

Another farfetched, but interesting application would be to identify a terrorist or criminal network based on the activities of a single person of interest. This would however require active monitoring the location data of every person in an area in real-time which would raise arguments about privacy and ethics.

Our datasets were largely collected through voluntary participants checking in through a simple web based interface. We collected data for over four months before we began analyzing the data. We tested the dataset with a co-location algorithm with and without considering the significance of the location of co-location which showed a significant improvement in the accuracy for the latter. Based on the results of this experiment we propose a system to identify LSGs in social networks. While testing the LSG generationalgorithm, we conducted a survey to test the accuracy of our prediction and found that it was 60-95% accurate in identifying social groups.

The rest of the paper is organized as follows: Section 2 describes the problem definition. Section 3 explains the GEOSO [9] model, and our modifications to it. Also Section 3 contains the LSG discovery framework and results of our experiments. The work is concluded in Section 4.

2. PROBLEM DEFINITION
Let D = (P1, P2 … Pn) be a set of mobility data collected from M users. Here Vi is a triplet of the form,

Pk = <user, location, time>

Pk is a record that tells when a user visited a particular location and when. Pk is only considered if the user has stayed in that particular location for a critical amount of time t. These locations are called staypoints [8]. This prevents erroneous

values from getting into the system as most activities that result in a relationship tend to happen at a staypoints rather than in transit. Even though transit points may be significant at a certain level, integrating that into the system is beyond the scope of this research.

Having D, we are interested in identifying and measuring the strength of the relationship between two users i and j. We are further interested in identifying a set of users G who are strongly related to each other.

3. DATA

Location data is available across the internet through various sources. The easiest source of location data are through Location Based Social Networks (LBSN) like Foursqaure. These LBSNs uses a check-in system where the user could check into a particular location. These check-ins are perhaps the most direct source of location of data for the purpose of our paper.

Another source of location data is through location tagged posts on Facebook and Twitter. Facebook and Twitter lets users tag posts with the location. Also Facebook images, are automatically tagged with the location extracted from the images EXIF data. Also the time of the status update can also be extracted to a rough approximation. This data can be extracted if the users give sufficient permissions to use their data. A third source of information is Flickr. Many studies have explored the use of Flickr for location data of users [4]. Flickr photos taken with cameras containing a GPS module, has the location as well as time of the photo in its EXIF data. Also, comments and user description of images provide location information.

Another approach towards finding the location of a user would be to leverage cellular network providers. This scenario however is quite problematic considering the privacy and ethical issues surrounding mass surveillance. However government agencies can utilize this system provided the privacy of the users are protected. This would involve removing personal identifiers from the location data. Cellular network providers can track the location of a user by simply triangulating the cell phones signal. Several smartphone operating systems also contain systems to obtain the real time position of the user continuously using various location services available in today's smartphones.

For the purpose of our research we employed a simple web application using the geoPosition JavaScript library. The geoPosition library smooths over the difference between the geolocation API, IP geolocation services, and the APIs provided by mobile platforms. The application enables users to check in their location anytime anywhere in the world. This provides us with an easy source of data for our research.

4. DISCOVERING SOCIAL TIES

4.1 Generating Visit and Co-Occurrence Vectors

The initial step in preparing the data using the GEOSO model is to create a visit vector for each user. Each dimension of the visit vector represents a unique geographic location or cell. For the purposes of our research we will divide the earth into cells of uniform size and associate an identifier with each. Now, the visit vector will store the time when the user visited each cell. If the user has not visited a cell during the observation time, the value associated with that cell will be zero. A visit vector for a user *i* will be of the form:

Vi = (cell_idx, <visit times>, …)

Here, visit times may have more than one value representing repeated visits to the same location. Visit times can either be timestamps or unique identifier identifying an interval of time.

Next a co-occurrence vector records the number of times two users have been together at the same time at the same place. Each dimension still represents a cell in the grid. The co-occurrence vector is represented as follows:

$C_{ij} = (c_{i1,j1}, c_{i2,j2}, \ldots, c_{iN,jN})$

The co-occurrence vector forms the basis of our remaining analysis.

4.2 Co-occurrence Location Significnace

Even though colocation .is a good indicator of a social relationship, it however fails in certain social scenarios. For example people working in the same building would show strong relationships in our system since this would generate a large number of co-occurrences. Similar situations can arise from colleges, schools, public transport, etc. In order to solve this issue, we propose a Location Significance Factor (LSF) for each cell in the grid. Each cell in the grid will be assigned the location significance factor based on how relevant a colocation at that location would be. For example, a coffee shop would have a high LSF factor while a location where people tend to gather naturally would have a significantly lower LSF factor. The LSF factor of a cell is an approximate probability that a colocation at that cell alone would be the result of two people having a relationship. For the purpose of our experiment we used a small LSF table based on our knowledge of the geographical area. However a better approach would be to calculate the LSF factor based on a known social network and mobility data.

The next step is to generate a modified co-occurrence vector by applying the LSF factor. This is done by simply multiplying the values in the co-occurrence vector with the corresponding LSF factor. The co-occurrence vector Cij now becomes

$Cij = (ci1,j1 * LSF1, ci2,j2 * LSF2, \ldots, ciN,jN * LSFN)$

The modified co-occurrence vector now should yield a better result when the relationships are estimated.

4.3 Normalizing Co-Occurrence Vector

Co-occurrence vectors can vary considerably among each other. Hence we normalize the vector by generating a normal vector which contains the maximum value of each cell among set of users. The normalizing vector N is defined as:

N=(m1,m2,m3…mn) mk = max(Cjk,jk)

4.4 Determining the strength of a relationship

The strength of the relationship is found by finding the pure Euclidean Distance between our modified co-occurrence vector and the Normalizing Vector N.

$$d_{ij} = \sqrt{\sum_k (c_{ik,jk} - m_k)^2}$$

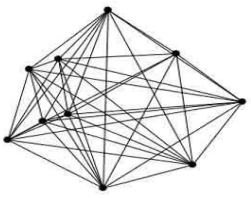

Fig 1: Social Network generated using the modified GEOSO model

For a group of users G, we generate the strength of relationship between each pair of users. This data can now be mapped into a weighted social graph of people. Each node in the graph is an individual and each edge represents the strength of the relationship between the user and another user. This is stored in an adjacency matrix for our use in the group discovery part of our research.

5. LSG DISCOVERY

Once the strength of social relationships are determined over a fixed geographic area, we can begin to identify local social groups within this graph. This is accomplished through a straight forward Markov graph clustering algorithm [10] on the weighted graph.

A simple clustering operation would yield sizeable clusters in geographic locations. Care must be taken in choosing the required parameters for the clustering so as to identify only clusters above a certain thresh-hold.

An important application of detecting online social groups is to detect place focused social groups. Also as the number of people involved in the group increases, the number of relationships to be calculated also increases exponentially. That is, if the system was considering the entire earth as the grid superset, the system would fail due to its exponential complexity. So we came up with an approach to identify LSGs in small geographic subsets, so as to enable the most practical use of our model. To discover LSGs, the size of the grid to be chosen can be a city or even a part of the city. However, if the user travels often to other parts of the grid, our model would again fail. In order to address this user's multiple social networks in different grids are stored as a subclass of his social network.

This makes it easier to use the data in practical applications. The steps followed to perform this operation can be illustrated as below:

For target user ut, identify his home grid GH.

1. Identify other users in current grid G.
2. Measure relationships for all other users in GH with user and generated a social network graph St for user ut.
3. Identify user's LSG and store information as home LSG for user.
4. If a user checks into a grid besides his home grid, perform steps 2 and 3.
5. Add new graph as a new LSG for user.
6. Repeat and store all sub clusters.
7. Recompute results periodically to ensure no new nodes have entered the system.
8. If result has to be retrieved, traverse the solution tree to obtain the corresponding result.
9. If the result to be retrieved is a larger subset of a set of grids, perform agglomerative clustering to generate the result set.

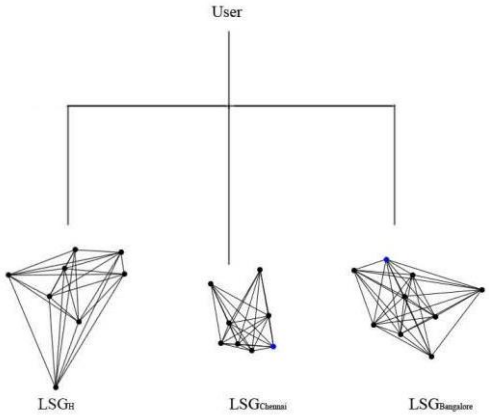

Fig 2: Representation of the subcluster representation

Storing a user's social relationships as many smaller LSGs provides for better computational as well as space complexity since the number of pairs of users are reduced over a smaller grid. Also retrieving a user's LSG would also be considerably easier when stored in a hierarchical structure.

6. RESULTS

6.1 Dataset

We now evaluate the effectiveness of our system to detect social groups in a local area by using their mobility data. For the experiment we handpicked 50 users of which some had relationships with each other of various degrees. The subjects were given access to our app and asked to check-in periodically. Next the users were asked to provide permissions to our Facebook app so that we could retrieve their Facebook friend list. Also, we extracted posts from the user, and identified location tags associated with them and added them to our dataset.

6.2 Approach

We ran our analysis on this dataset continuously for a period of three months after which the results were analyzed. Our first approach was to use the normal GEOSO approach which gave us a network graph of users and their relationships. Most of the 50 subjects studied in a common college. When the normal GEOSO locations were analyzed the number was co-occurrences was quite high as the students were at the college every day and checked in multiple times from the same location. The results of the GEOSO distance measure also showed a high degree of relationships among all subjects who studied in the same college.

Our next approach was to run our modified GEOSO relationship metric system to analyze the strength of the relationship among the same people. In our modified GEOSO approach we used an LSF factor of 0 for the college that was the most common colocation factor which did not significantly aid the relationship between people.

Now, once we computed the relationship between people, we obtained a social network graph that was more natural. This graph contained more number of clusters and was not as tightly bound as the result of the previous experiment.

Once we had the social graphs of the users our next approach was to identify LSGs among these graphs. We ran the solution through a Markov clustering model which gave us multiple clusters within the relationship graphs. After we generated the initial set of LSGs, we verified the accuracy by asking each users to identify the people in their respective LSGs, the people they did not personally know or did not have a close relationship

with. This was to be our metric to verify the accuracy of our approach. We asked them to perform this verification on the initial normal GEOSO results as well as on the modified GEOSO results. The results of the same can be illustrated in Fig 3.

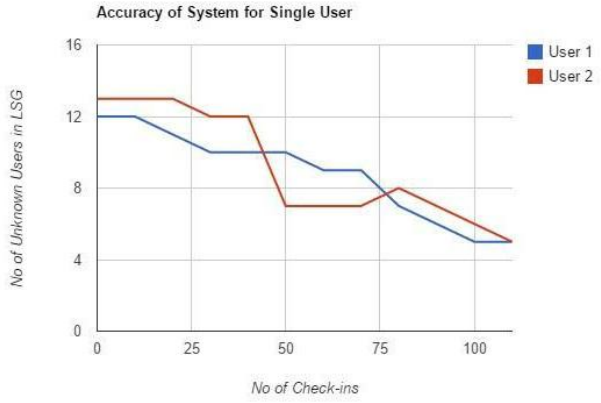

Fig 3: Accuracy of LSG vs No of Check-ins using normal GEOSO

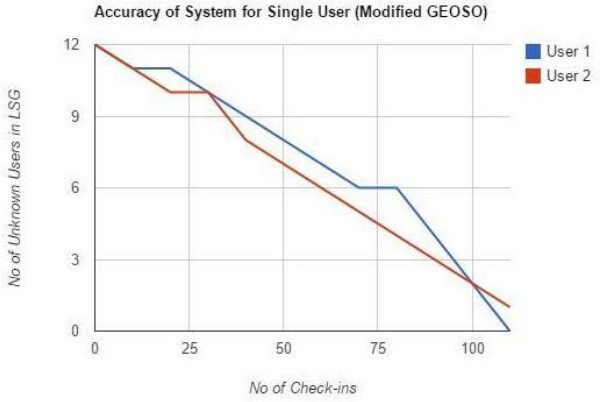

Fig 4: Accuracy of LSG vs No of Check-ins using modified GEOSO

It is also interesting to note that the number of check-ins or the number of data points available with us is directly related to the accuracy of our system. We can observe an almost direct proportionality to the accuracy of the system. This is due to the fact that more data points allow for more number of colocations to be detected and hence improves the efficiency of our system.

6.3 Effect of LSF on Results

The effect of the LSF on the results was interesting observation in our experiment. To test the effectiveness of our observation, we used the Google Place Types [11] to identify the type of place the user checked into. Now we applied a LSF table as shown in Table 1 to generate a modified colocation vector.

Table I: LSF Values for Commonplace Types

Type of Place	LSF
Bar	0.90
Coffee-Shop	0.95
Bank	0.3
Bowling Alley	0.85
University	0.1
Establishment	0.1
Hospital	0.15
Bus Station	0.1
Movie Theater	0.70
Temple/Mosque/Church	0.5
Zoo	0.8
Night Club	0.65
Restaurant	0.7
Stadium	0.4

This LSF table was generated only for the locations available in our data. Now when we compare the results of the previous relationships to the new one we can observe from the graph the differences in cluster formation. Through the normal GEOSO metric the graph tends to be very closely knit together with the number of elements almost equal to the number of students going to the same university. However in our second approach when we muted the effect of the colocations at the college we obtained a significantly better, natural looking graph shape with multiple smaller clusters among the users.

6.4 Effect of Size of Grid

The size of the grids that we consider also affects the accuracy of our system. In order to study this, we considered a small area around the university where the subjects studied as a grid and ran our simulation for co-occurrences. After our simulation we observed that the size of the grid does affect the accuracy of the grid. A smaller grid ensured that co-occurrences that occurred were due to the relationship that existed between the people. The LSGs that were identified in the smaller grid were merely a subset of the original social network identified over the entire city. However the number of unknown users in the LSG was much lesser compared to the previous results.

However when the size of the grid was small, the number of people in the LSG was significantly smaller compared to the number of users found in the previous system. This means that the LSGs also missed out on some relationships that were detected in the previous approach. Selecting the grid size is the choice of the user and further research is required in detecting an optimal size for the area of the grid.

Table II: Accuracy for Single User in Various Grid Sizes

Grid	No of Check-ins	No. of unknown users in LSG	No. of missed connections
Full City	234	17	2
Area 1	65	4	22
Area 2	43	8	12

7. FUTURE WORK

These results leave show a lot of possibility in the use of mobility data to determine the effect of relationships among people. The first area that could be focused in future works is the calculation of an optimal LSF value that can better describe the social relationships among people. A probabilistic model must be utilized to setup a common LSF index for all locations in the area of study.

A second direction of future improvement is in the cluster detection realm of social networks. Social networks are peculiar in terms of their graph structure, which reflects the complexities of human social behavior. This paper used Markov Graph

Clustering, however we feel that better algorithms need to be developed to improve the detection of social groups within social networks in both the online as well as offline realms.

Another area that can be improved further is the data sources. Spatiotemporal data sources are plenty. Research into a common framework for the extraction of spatiotemporal data from multiple social networks is very much needed. Such a system should be able to extract spatiotemporal information from popular social networking services like Facebook and Twitter, and also from other sources including mobile applications and Location Based Social Networks. This would provide a big data set for research into the human sociological behavior similar to our research.

8. CONCLUSION

In this paper we proposed a model to extract the social relationships of people using mobility data. We use a modified GeoSocial model to quantify the strength of relationships.

We used a Location Significance Factor to identify the significance of the location of co-occurrence. This enables the system to eliminate co-occurrences that were not the result of social relationships. This provided us a way to better the existing GEOSO system and improve the similarity measures of the system drastically.

We also used devised a method to detect social groups within social networks and also an efficient way to store it by creating multiple instances of a person's social network over smaller geographic areas and aggregating them to generate the person's complete social circle.

In our experiments, we were able to extract better results than the existing GEOSO model and was also able to examine our proposed model by examining the relationships of people with a known social relationships and verifying the results with the users themselves. Our system showed significant efficiency in detecting social groups with upwards of 65% percent accuracy. The accuracy further depended on the various factors and while examining multiple scenarios we observed a maximum accuracy of 95%.

However our dataset was small and we believe that a larger dataset will generate more complex scenarios owing to the complex nature of human behavior. Therefore our future works will look forward to validating our ideas to a much larger dataset.

9. REFERENCES

[1] Foursquare. Available: https://foursquare.com/, Accessed on 25 April 2015

[2] Facebook. Available: https://www.facebook.com/, Accessed on 26 April 2015

[3] Twitter. Available: https://www.twitter.com/, Accessed on 25 April 2015.

[4] Claudia Hauff and Geert-Jan Houben, "Geo-Location Estimation of Flickr Images: Social Web Based Enrichment", 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings

[5] Flickr. Available: https://www.flickr.com, Accessed on 1 May 2015.

[6] Sriram Sankar, "Under the Hood: Indexing and ranking in Graph Search", Facebook Engineering

[7] Scott L Feld, "The Focused Organization of Social Ties", American Journal of Sociology, Vol. 86, No. 5 (Mar., 1981), pp. 1015-1035

[8] Mitra Baratchi, Nirvana Meratnia, Paul.J.M. Havinga, "On the Use of Mobility Data for Discovery and Description of Social Ties", Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 25-28 Aug 2013, Niagara falls, Canada. pp. 1229-1236

[9] Huy Pham, Ling Hu, Cyrus Shahabi, "GEOSO – A Geo-Social Model: From Real-World Co-occurrences to Social Connections",

Databases in Networked Information Systems: 7th International Workshop

[10] Van Dongen, S, "Graph Clustering by Flow Simulation", PhD Thesis, University of Utrecht, The Netherlands.

[11] Google Place Types. Available: https://developers.google.com/places/supported_types,

[12] Chloë Brown, Vincenzo Nicosia, Salvatore Scellato, Anastasios Noulas, Cecilia Mascolo, "The Importance of Being Placefriends: Discovering Location-focused Online Communities", Proceedings of the 2012 ACM workshop on Workshop on online social networks, Pages 31-36

[13] David J. Crandalla, Lars Backstromb, Dan Cosleyc, Siddharth Surib, Daniel Huttenlocherb, and Jon Kleinbergb, "The importance of being placefriends: discovering location-focused online communities", Proceedings of the 2012 ACM workshop on Workshop on online social networks, Pages 31-36

[14] Eunjoon Cho, Seth A. Myers, Jure Leskovec, "Friendship and mobility: user movement in location-based social networks", Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1082-1090

[15] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, Albert-László Barabási, "Human Mobility, Social Ties, and Link Prediction", Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1100-1108