Community Detection in Cellular Network Traces

Mariya Zheleva, Paul Schmitt, Morgan Vigil and Elizabeth Belding

Department of Computer Science
University of California, Santa Barbara
{mariya, pschmitt, mvigil, ebelding} @cs.ucsb.edu

ABSTRACT

Studies of user behavior in cellular networks have served as a knowledge base for development of critical applications and services catered to specific user needs. In this paper we examine community persistence in egocentric social graphs extracted from cellular network traces in the Cote d'Ivoire provided by Orange. The goal of our study is to inform mechanisms for improved dissemination of information by identifying subscribers or groups that can serve as information relays. We find that communities that persist in an egocentric network are independent of one another. Thus, multiple information relays can be selected from each independent community, to increase the probability that information will flow to the ego.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: Network Operations; C.4 **[Performance of Systems]:** Modeling Techniques; E.1 **[Data Structures]:** Graphs and Networks

General Terms

Algorithms, Measurement, Human Factors.

Keywords

Cellular networks, egocentric graphs, community detection, temporal graph mining.

1 INTRODUCTION

The availability of mobile networks has revolutionized the way people communicate in the developing world. Our first hand experience in rural Macha, Zambia indicates that access to cellular services is of critical importance to residents. While the reasons for adoption of cellphone technology in developing communities are not drastically different than those of the Western world, the benefits for people in these remote communities without infrastructure or other means of telecommunications is much more pronounced. Obtaining information via cell phone, as opposed to in person after travel, saves both critical time and money.

A plethora of applications that improve the well-being of people in remote communities leverage cellular networks. Such applications span from health care [3] and education [1] to agriculture [11] and mobile banking [9]. Multiple successful projects in Africa have originated from observing user behavior in mobile or social networks. As a result of Facebook traffic analysis, Johnson et al. designed a system to facilitate local content sharing within remote rural communities [7]. Mbiti et al. describe a system called mPesa [9] that enables transfer of money in the form of airtime in rural Kenya. The design of this system was inspired by analysis of mobile network usage in Kenya, which indicates that people tend to transfer airtime between one another as a means for payment or financial support. Follow up studies on the adoption of mPesa in Kenya show that theft decreased, as users no longer needed to carry cash.

Such projects are of critical importance to introducing new services and enhancing the well-being of people in under-serviced areas. At the same time, special attention should be paid in the design process of these systems to make sure that they meet an actual need in the community. Analysis of large scale datasets generated by the targeted communities naturally facilitates the identification of actual community needs.

We approach a cellular network dataset from Cote d'Ivoire with this end in mind. The dataset provides information for the personal network of 5,000 randomly selected individuals; these personal networks are called egocentric social graphs. We analyze these egocentric social graphs hoping to identify community persistence in an attempt to motivate feasibility of information relays in user-centered cellular communication.

Social network analyses using mobile traces focus on implications of network diversity [5], extracting relations [6] and community formation [10]. These works, however, are not concerned with temporal aspects of individuals' communication networks. This paper makes several contributions. First, we design a model based on persistence graphs to study temporal persistence of social groups in egocentric graphs. We then discover that while there is a weak community persistence in egocentric graphs, there are individuals in an egocentric network that are highly persistent over time.

2 METHODOLOGY

We analyze individual user communication patterns over time. In particular, we look at community persistence in egocentric social graphs, whereby a subscriber of interest is centered in a graph and the periphery nodes of this graph are other subscribers with whom the central node communicates. In this section we start by describing the dataset as provided by Orange. We then describe our model for egocentric social graph analysis.

2.1 Dataset

Our analysis is based on a dataset that features the personal communication networks of 5,000 subscribers (egos) and was provided by Orange. The dataset was collected in Cote d'Ivoire over the course of 150 days between December 1, 2011 and April 28, 2012. To assure homogeneity, the data includes records only for users who were subscribed with the network for the entire capture period. In this dataset the capture period is divided into

ICTD'13, December 7-10, 2013, Cape Town, South Africa.

http://dx.doi.org/10.1145/2517899.2517932

ten equal sub-periods, each of which contains one network per ego. This subdivision of data in time allows temporal analysis of individual communication networks. These personal communication networks include up to second degree neighbors of an individual and are called egocentric graphs. An edge between neighbors in these ego-graphs indicates that there was at least one call between the two users; no information for number of calls, call duration or direction is provided. Edges are drawn between (i) the ego and its first order neighbors, (ii) between two first order neighbors or (iii) between first and second order neighbors.

The 5,000 egocentric graphs were anonymized by Orange before release to researchers. The process of anonymization kept the ego as well as the neighbor ID the same throughout the entire observed period. If the same subscriber, however, appeared in more than one egocentric graph, the subscriber's ID is different in the different graphs. While the telecom did not provide information regarding the randomization process, the dataset is considered representative of the user population in Cote d'Ivoire.

2.2 Egocentric Graphs Analysis

We examine the egocentric social graphs dataset to determine persistence of social groups for each ego over time. We also analyze the likelihood that one or more nodes (users with which an ego communicates) persist over time in an egocentric graph. We hope to see persistence in both communities as well as individual subscribers. We hypothesize that such continuously-present entities can be used as information relays to strengthen information distribution amongst community members. Our analysis indicates that while community persistence is relatively low, persistent nodes indeed exist.

In order to extract the separate social groups of an ego, we remove the ego node from each egocentric social graph (Figure 1) and analyze the connected components that remain. Each connected component corresponds to one social group. Note that in the text we use the terms connected component and social group interchangeably.

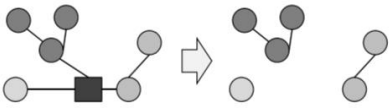

Figure 1: The effect of removing the ego (depicted with a square) from the egocentric social graph

After extracting the connected components we evaluate the persistence of these components over time. A connected component is 100% persistent over two consecutive periods if the nodes in this connected component are identical in the two periods. For this evaluation we define a persistence graph G=(N,E,W) with N nodes, E edges and W weights assigned to each edge. Each node in G is a connected component labeled with the period to which it belongs. An edge exists between two connected components if they overlap in consecutive periods. The weight assigned to each edge is the Jaccard similarity, J, between the connected components [12]. For two sets A and B, the Jaccard similarity J can be calculated as follows:

$$J = \frac{A \cap B}{A \cup B} \quad (1)$$

J ranges between 0 and 1, where 0 indicates no overlap and 1 indicates full overlap.

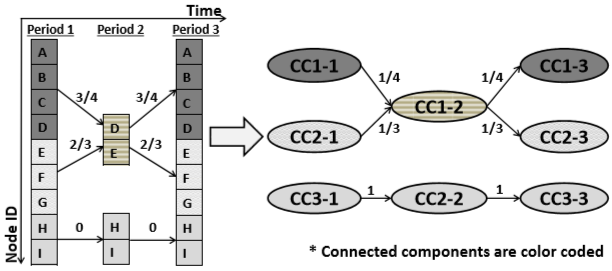

Figure 2: Building a persistence graph.

Figure 2 presents an example of building the persistence graph for a single ego over three consecutive periods. The left side of the picture presents the set of neighbors in each of the three periods. The social groups comprised by these neighbors are color-coded. The right side of the picture presents the resulting persistence graph. Each node corresponds to a connected component (CC) in a given period. In the figure node labels are of the format CCID-PeriodID. Edges exist only between connected components that overlap fully or partially in consecutive periods. There is no edge between connected components that persist over non-consecutive periods (e.g. there is no edge between node "CC1-1" and node "CC1-3").

Our persistence analysis is based on the described persistence graphs and consists of two parts. First, we analyze the in- and out-degree distribution of the nodes in the persistence graph. We note that if the social groups of an ego persist over time, all the nodes in the persistence graph should have in- and out-degrees of either 0, if the node belongs to the first or last period, or 1, if the node is in the intermediate periods. In cases where social groups do not persist, nodes can have a degree of 0 if the corresponding social group does not re-appear in following periods. Nodes can also have in- and out-degrees larger than 1 if social groups merge or split in consecutive periods.

We attempt to quantify the level to which social groups overlap by considering the weights of the edges in the persistence graphs. As detailed earlier, edges are drawn between nodes that overlap fully or partially in consecutive time periods. The weights assigned to these edges are the Jaccard similarity between the nodes connected by these edges. For each transition between period *t* and period *t + 1* we find the normalized Jaccard similarity $JS^{(t,t+1)}$ between these periods; that is the sum of edge weights $W_i^{(t,t+1)}$ divided by the number of edges $|E^{(t,t+1)}|$ between the two periods:

$$\hat{JS}^{(t,t+1)} = \frac{\sum_{i=1}^{|E^{(t,t+1)}|} W_i^{(t,t+1)}}{|E^{(t,t+1)}|} \quad (2)$$

We then find the average Jaccard similarity for the entire persistence graph by summing the normalized Jaccard similarities and dividing this sum by the number of period transitions K.

$$\bar{JS} = \frac{\sum_{j=1}^{K} \hat{JS}_j^{(t,t+1)}}{K} \quad (3)$$

Informally, the higher the average Jaccard similarity, the more persistent the social graphs of an ego are over time.

3 ANALYSIS RESULTS

We start our analysis with evaluation of the average number of social groups with which each ego communicates over the entire capture period from December 2011 to April 2012. For this analysis we sum the number of connected components that appear in each two-week period and divide this sum by the number of capture periods. Figure 3(a) plots a CDF of the average number of connected components for each ego. While the average number of components across egos spans from 1 to 10, 68% of egos have between 2 and 5 connected components on average. Further, we examine how the number of connected components deviates for each ego. Figure 3(b) plots a CDF of the standard deviation of the number of connected components per ego over the observed period. Almost half of the egos (47%) have standard deviation of less than 1, while 96% of all the egos have standard deviation of less than 4. This indicates that the number of connected components in an egocentric graph remains relatively constant over time.

Next we analyze the persistence of these social groups over time. First, we look at the in- and out-degree distribution of nodes in the persistence graphs. As detailed in Section 2.2, a node in period *t* has in- or out-degree of 0 if it belongs to the first or last observed period or if it does not overlap with any node from the preceding $t - 1$ or the following $t + 1$ period. Nodes have in- and out-degree of exactly 1 if they persist over time, and degree larger than 1 if they split or merge over consecutive periods.

We calculate that out of all the nodes in all persistence graphs, 9.49% belong to the first period (i.e. have in-degree of 0) and 8.93% belong to the last period (i.e. have out- degree of 0). At the same time Figure 4(a) indicates that in nearly 60% of the cases nodes have in- or out-degree of 0. This means that about 50% of all the social groups that we observe, and which were not associated with the first or last period, did not occur in the preceding and following periods. 40% of the nodes have in- or out-degree of 1, indicating that 40% of the social groups persisted in consecutive periods. Only about 3% of the cases have in- or out-degree larger than 1; social groups rarely split or merge over consecutive periods.

This result indicates an important quality of the observed egocentric social graphs: there are two distinctive types of social groups with which an ego communicates – (i) those that likely occur only once (in- and out-degree is 0), and (ii) those that likely persist over time and strictly correspond to one social group from the preceding and one social group from the following period.

The former group can be associated with one-time calls, for example calling to schedule a doctor appointment, while the latter can be associated with calls recurring over time, such as these between relatives and friends who stay in touch.

We continue our evaluation of social group persistence by analyzing the weight of edges (representing the similarity) of social groups in consecutive periods. We leverage the average Jaccard similarity metric as defined in Section 2.2; the closer this similarity is to 1, the larger the overlap between social groups in consecutive periods. Figure 4(b) plots a CDF of the average Jaccard similarity for the 5,000 egocentric graphs. The median of this CDF is only 0.22, which means that on average the overlap of social groups over time is relatively small – about 22%.

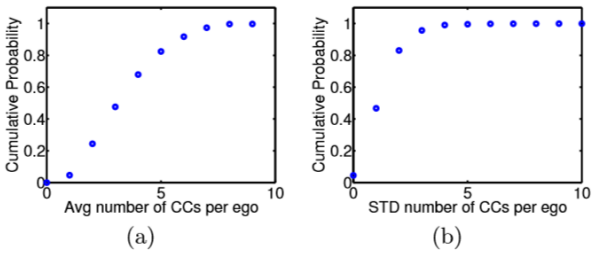

Figure 3: (a) The number of connected components (CCs) per ego and (b) the standard deviation of the number of connected components per ego over the observed period

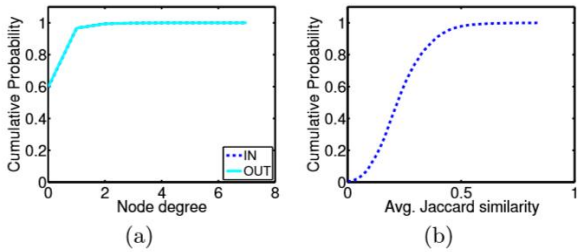

Figure 4: (a) The in- and out-degree of nodes in all persistence graphs and (b) the average Jaccard similarity for each persistence graph

Finally, we evaluate the frequency of occurrence of the neighbor that appears most often in the social network of an ego. For this evaluation we count in how many of the ten observed periods each neighbor appears. We then sort the neighbors in decreasing order of appearance frequency. We compare the first, second and tenth most frequent neighbors to determine whether there are groups of neighbors that appear more often and what the typical size is of such groups.

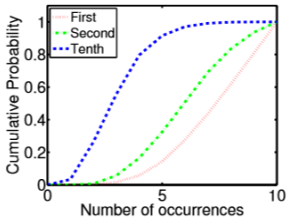

Figure 5: Number of occurrences of the first, second and tenth most frequent neighboior

Figure 5 presents our results. The median value for the first top neighbor is 8, while for the second and the tenth top neighbor it decreases to 6 and 3, respectively. This means that in 50% of the cases, the most frequently occurring neighbor exists in 8 out of the 10 observed periods. These results indicate high persistence of at least one neighbor in the social graph. At the same time, a group of two most persistent neighbors would appear ten times in only 6.8% of the cases, which indicates that a group of most persistent neighbors would typically have very few members.

4 RELATED WORK

Analysis of mobile network traces provides a unique opportunity for large-scale verification of socio-economic models that were previously derived and studied on much smaller scales. Previous research related to our work can be divided in two groups: (i) social interactions analysis and (ii) dynamic graph mining.

Social analysis. Social network analyses based on cellular traces focus on implications of network diversity [5], extracting relations [6] and community formation [10]. Studies demonstrate that diversity of one's mobile social network influences socio-economical prosperity [5]. Other work extracts biases in self-reported friendships by comparing characteristics of self-reported relationships with those extracted from cellular traces [6]. These studies, however, are not concerned with variability of social networks over time. In contrast our analysis explores temporal trends of cellular communication in individual subscribers' communication networks and provides insights on community persistence in egocentric social graphs.

Dynamic graph mining. In the area of dynamic graph mining, research has focused on evolutionary community detection [8], conserved relational states [2] and high-scoring dynamic subgraphs [4]. Bogdanov et al. propose a method to identify the highest-scoring temporal subgraph (e.g. most congested road segment) in a dynamic network [4]. Our analysis is different, as we seek to summarize the persistence in different egocentric networks without observing the whole graph at a time. Other work mines relational patterns in a dynamic network [2] in order to detect maximal evolution paths in time-evolving networks. While this work utilizes a model for tracking similarity that is similar to our persistence graphs, the proposed scheme is only concerned with full overlap of graph entities over time. In contrast, our method captures partial overlaps and allows for fine-grained analysis of community persistence.

5 DISCUSSION AND CONCLUSION

We present preliminary analysis on community persistence in an egocentric network. Analyzing 5,000 random users from Cote d'Ivoire, we find that on average an egocentric network has four social groups; this number is stable over time. We also find that 50% of the observed social groups did not occur in the corresponding preceding and following periods. At the same time, less than 1% of the communities split or merge over time; thus communities that do persist tend to be independent of one another. Finally, we observe that on average there is a 22% overlap of social groups over time. Persistence of a subscriber in one's social group likely means that there is a stronger personal connection between the ego and the corresponding subscriber. We hope that with this preliminary analysis we can inform mechanisms for improved communication channels in rural communities by the use of information relays in egocentric networks. Our first-hand experience in rural health care indicates that improvement of information channels is of critical importance since health care services such as vaccinations are often available; however, it is difficult to bring information about availability to the interested patients. While our analysis provides some insights on community persistence, in order to devise models for information relay extraction, more detailed data is needed that contains information such as location, frequency, duration and type of interaction. Availability of such information will enable true extraction of individuals that are strongly connected to an ego and can serve as reliable information relays.

6 ACKNOWLEDGEMENTS

We thank Orange for providing data to facilitate our analysis. This work was supported in part through NSF Network Science and Engineering (NetSE) award CNS-1064821.

7 REFERENCES

[1] Dr Math – remote math tutoring using MXIT in South Africa. http://www.elearning-africa.com/eLA_Newsportal/mixing-it-with-dr-math-mobile-tutoring-on-demand/ .

[2] R. Ahmed and G. Karypis. Algorithms for mining the evolution of conserved relational states in dynamic networks. In ICDM, 2011.

[3] R. Anderson, E. Blantz, D. Lubinski, E. O'Rourke, M. Summer, and K. Yousoufian. Smart connect: last mile data connectivity for rural health facilities. In NSDR, San Francisco, CA, 2010.

[4] P. Bogdanov, M. Mongiovi, and A. Singh. Mining heavy subgraphs in time-evolving networks. In ICDM, 2011.

[5] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. In Science 21 May 2010: Vol. 328 no. 5981 pp. 1029-1031, May 2010.

[6] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. In PNAS, Vol. 106, No. 36., 2009.

[7] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. VillageShare: Facilitating content generation and sharing in rural networks. In Proceedings of the 2nd ACM Symposium on Computing for Development, ACM DEV '12, pages 7:1–7:10, New York, NY, USA, 2012. ACM.

[8] M.-S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. Proc. VLDB Endow., 2(1):622–633, Aug. 2009.

[9] I. Mbiti and D. N. Weil. Mobile Banking: The Impact of M-Pesa in Kenya. Working Paper 17129, National Bureau of Economic Research, June 2011.

[10] J. Onnela, S. Arbesman, M. Gonzalez, A. Barabasi, and N. Christakis. Geographic constraints on social network groups. PLoS ONE, 6(4):e16939, 04 2011.

[11] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural India. In CHI, Atlanta, GA, 2010.

[12] P.-N. Tan, M. Steinbach, and V. Kumar, editors. Introduction to Data Mining. Addison Wesley, 2005.