Understanding communities using mobile network big data

CPR*south* 2015

Kaushalya Madhawa, Sriganesh Lokanathan, Rohan Samarajiva, Danaja Maldeniya

July 2015

LIRNE*asia*
info@lirneasia.net | www.lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160.
info@lirneasia.net
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada and the Department for International Development (DFID), UK.

Abstract

The patterns of human interactions are not random, but intertwined with many attributes of individuals such as ethnicity, economic status etc. Using anonymized call detail records obtained form a mobile operator in Sri Lanka, this paper investigates the communities formed by the communication patterns. By applying several community detection algorithms, we could identify a community structure consisting of 11 communities as the most suitable one. These communities show similarity to the nine provincial boundaries at varying degrees. But all these communities show high level of spatial coherence. Additionally we explore how these communities segment into a further level of sub-communities.

Keywords: Mobile Network Big Data, Call Detail Records, Social Geography, Community Boundaries

1 Introduction

Recent advancements in computing power and access to large-scale communication data have had major implications for the social sciences. This has enabled the study of human geography especially on human relationships and societal structures, at a scale that was not possible before. The near-ubiquitous use of mobile phones and the resultant digital spatio-temporal footprints left by the population as they travel and communicate affords a quantum improvement in terms of base data to study human geography.

Understanding the strength of social connections and identifying communities amongst the population is valuable for modeling disease spread, information flow, and mobility patterns. Administrative boundaries, formed by history and geography, do not necessarily reflect the actual communities or social interaction patterns within a region. These often depend on shared attributes of individuals in a region, such as gender, ethnicity, economic status, etc. These shared attributes are often reflected in the communication activity.

Having negotiated access to communication data from an operator in Sri Lanka, we leverage the recent advancements in social network theory to understand community structures in Sri Lanka. Having emerged from a civil war in 2009 and with the underlying social, ethnic and religious tensions yet to be fully resolved, the insights from this work can have broader socio-political implications, beyond just pure social science research.

2 The state of the art in community detection

Recent advancements in computing power and access to large-scale communication data have had major implications for the social sciences. This has enabled the study of human geography especially on societal structures and human relationships, at a scale that was not possible before. The findings of previous research (Onnela et al., 2007a; Calabrese, Smoreda, Blondel, & Ratti, 2011) suggest that insights gained from one communication medium (e.g. phone network) can be generalized to other human interaction networks (e.g. face-to-face) as well. This finding and the abundance of mobile phones have lead Mobile Network Big Data (MNBD) to be used as a proxy for human interactions.

Decomposing a social network into groups of densely interconnected nodes or communities (Fortunato & Castellano, 2012), can be used to uncover relationships which are not visible at the global level. The presence of strong ties among individuals lead to the formation of communities (Onnela et al., 2007b).

Application of community detection algorithms on mobile Call Detail Records (CDR) data in Belgium could detect the spatial split in different linguistic regions (Blondel and Guillaume, 2008). The resultant communities on spatially aggregated social networks found to be geographically cohesive. Similar work using landline phone calls in Great Britain (Ratti et al., 2010) and mobile CDR data in USA (Calebrese et al., 2011) found geographically cohesive communities that generally correspond to administrative boundaries.

In this study we employ community detection algorithms to a mobile CDR network in Sri Lanka to compare natural communities exist in the interaction network against administrative regions of Sri Lanka. Additionally we explore how these communities segment into a further level of sub-communities.

3 Data Source

The paper uses one month of Call Detail Records (CDRs) for nearly 5 million SIMs from an operator in Sri Lanka[1]. The data is completely pseudonymized by the operator i.e. the phone numbers have been replaced by a unique computer generated identifier. The researchers do not maintain any mapping information between the generated identifier and the original phone number.

Each CDR corresponds to a particular subscriber of the operator's network and is created every time a subscriber originates or receives a call. In the case of an in-network call (i.e. both parties on the call were subscribers in the same mobile network), two records are generated, one for each party. Each record contains the following attributes:

- Call direction: A code to denote if the record is an incoming or outgoing call
- Subscriber identifier: Anonymized identifier of subscriber in question
- Identifier of the other party: Anonymized identifier of the other party on the call
- Cell identifier: an ID of the cell (i.e. antenna) that the subscriber was connected to at the time of the call
- Date and time that the call was initiated
- Duration of the call

4 From digital traces to communities in Sri Lanka

We only considered in-network calls i.e. calls that originated and terminated on the same operator's network. In-network calls can be identified by virtue of the fact that there will be two records in the data for the same call (one for the incoming call and the other for the outgoing call). The resultant network formed by considering only the in-network calls can be considered a network of human interactions at the individual level. At this level the network

1 Due to the agreements with the operator we are unable to name the operator and cannot give a precise figure for the number of SIMs that were analyzed.

is too granular for our purposes. Therefore we aggregate the call network to the level of the Base Transceiver Station[2] (BTS), so that we end up with a network of BTS to BTS communication. This results in a graph consisting of 2549 nodes (i.e. BTSs). The total number of calls between two BTSs is used as an indicator of the strength of the interaction between that pair.

4.1 The influence of distance to communication

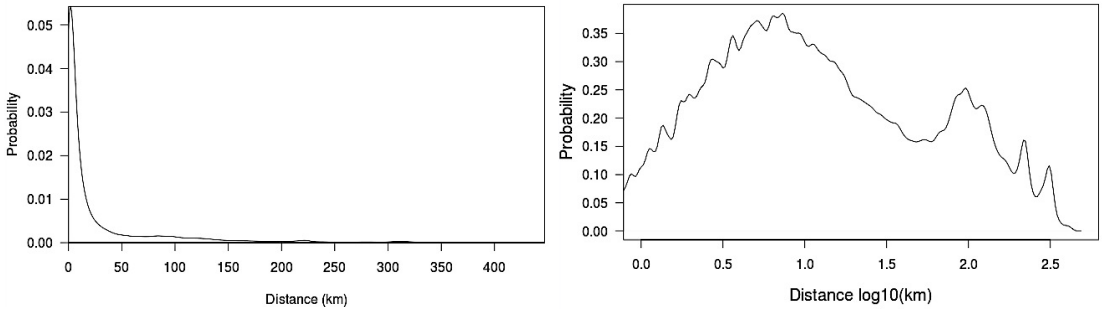


Figure 1: Distribution of the volume of inter-BTS communication by distance between BTSs

Figure 1 suggests that the call volume between two locations decays with the distance between them. Infact our findings mesh well with prior work by Krings, Calabrese, Ratti, & Blondel (2009), who using MNBD from Belgium, found that a power law fit well for call durations between cities more than 10km apart. We found a noticeable unexpected spike in call volumes when we consider the log distribution (Figure 1b). The sudden increase of call volumes is visible in the logarithm range of 2 to 2.3 (corresponding to inter BTS distances of between 100km and 200km). This cannot be explained by the power law in itself, and may be reflective of greater population in the areas covered by the respective base stations and/or higher economic activity in those areas. Colombo is Sri Lanka's commercial hub and most of the other major cities are located within the distance range 100 to 200km (e.g.: Kandy, Galle, Matara). This needs to be investigated further, but prima facie, this meshes with what we might expect from a gravity model, that has been used to explain diverse behavior from car traffic (Jung, Woo-Sung, Fengzhong, and Stanley, 2008) between two locations as well as trade flows (Tinbergen, 1962).

4.2 Communities

Even though the definition of a community can be relatively simple in terms of our common understanding, its underlying attributes (which can be different) lead to different formal mathematical definitions. These variations in the mathematical definitions will mean that the community structures under different mathematical formulations can vary greatly (Coscia, Giannotti, & Pedreschi, 2011). Detection of communities in large networks become computationally intractable as the size of the network grows. Hence a multitude of algorithms have formed around different mathematical formulations of what a community is or isn't. For example information theoretic algorithms (Rosvall & Bergstrom, 2008) aim to represent the network in a more denser structure. Modularity maximization algorithms (Newman, 2004) try to maximize modularity, a fitness function that defines the density of the internal edges in a cluster. Some other algorithms aim to uncover the community

2 A Base Transceiver Station (BTS) is the commonly utilized term for a base station. It consists of one ore more antennas and has a certain geographical coverage area.

structure using random walks (Pons & Latapy, 2005) along the edges. Rinzivillo, Mainardi, & Pezzoni, (2012) used information theoretic algorithm Infomap (Rosvall & Bergstrom, 2008) to detect communities in a network of GPS tracked vehicles in Pisa, Italy. They claim that this method is capable of detecting fine-grained communities compared to popular modularity maximization methods.

We decided to use modularity maximization techniques to understand Sri Lankan communities since we are able to do intra-community comparisons with just a single metric (i.e. modularity score). The modularity metric quantifies how dense the links within a community are as compared to links between communities. Modularity values range from -1 to 1 with higher values indicating clusters with greater intra-community connections than inter-community connections. The goal of modularity based community detection algorithms is to find a partitioning such that the overall average modularity is maximized. However it is computationally impossible to find the optimal partitioning for large networks. Therefore the extant modularity algorithms utilize approximation techniques to achieve a good modularity in a computationally feasible manner.

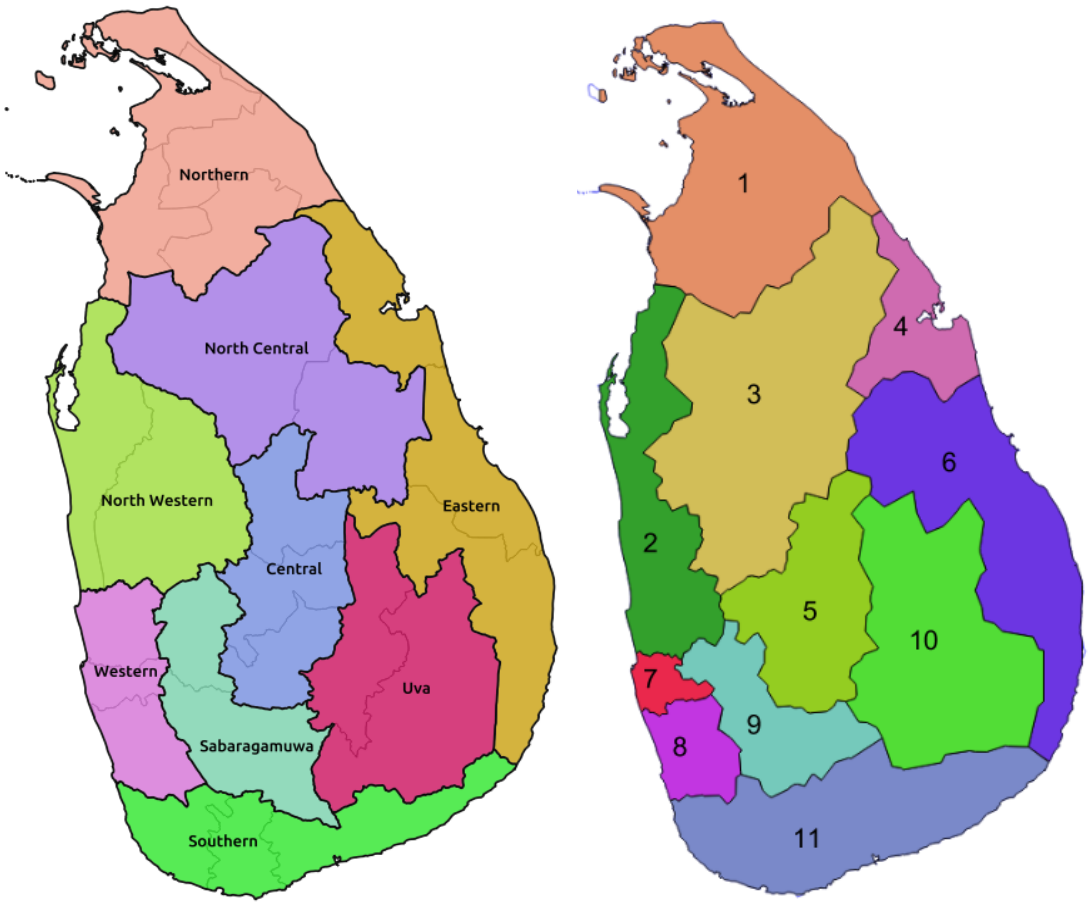
Previous work by others using a modularity optimization approach, have used different techniques. Blondel and Guillaume (2008) used a Fast-Greedy approach using data from Belgium. Sobolevsky et al (2013) applied a multilevel community detection technique (more commonly referred to as the Louvain method) when analyzing data from Portugal, Belgium, France, Italy, Saudi Arabia, Côte d'Ivoire, and the United Kingdom. In addition to these two techniques there are also others such as Walktrap (Pons & Latapy, 2005) and also Leading Eigenvector (Newman, 2006). But to our knowledge no one has so far applied the latter two techniques to MNBD. We applied all four techniques to our Sri Lanka data as shown in Table 1 and found the Louvain method to be the best, since it generated the highest average modularity (0.57).

Table 1: Comparison of multiple algorithms based on modularity

Algorithm	Modularity
Louvain	0.57
Fast-greedy	0.562
Walktrap	0.557
Leading eigenvector	0.520

The derived communities can be seen in Figure 2b. Even though no spatial assumptions were made in the method, the derived communities show a high degree of spatial coherence.

The primary administrative divisions in Sri Lanka are provinces, of which there are 9 in total. Each province is further divided in districts, with a total of 25 districts in Sri Lanka. Each district is further divided into Divisional Secretariat Divisions (DSDs) and there are a total of 331 DSDs in Sri Lanka. When comparing the map of the 9 provinces against the map of the 11 communities detected using modularity optimization technique on MNBD, we notice only a few similarities. Only the northern-most community (1) and the southern-most community (11) closely follow the existing boundaries of the Northern Province and Southern Province respectively. When it comes to differences there are many, the starkest of which is that there are 11 derived communities as compared to only 9.



(a) The 9 provinces of Sri Lanka **(b) The 11 communities detected from MNBD**

Figure 2: Comparison of provincial boundaries with the communities detected from MNBD

The Western Province, which actually consists of three districts, namely Gampaha, Colombo, and Kalutara, has broken off into 3 separate communities largely following the district boundaries with few notable differences. Most of Gampaha seems to have merged with the littoral areas of the North-Western province to form Community 4. Colombo district has become its own community (community 7) crossing the Kelaniya river and absorbing the southern parts of Gampaha, but shedding some of the north-eastern parts of the district. Colombo district is the largest in Sri Lanka by population (2.2 million people) and contains the economic capital of Colombo city, which make this district the core of the Sri Lankan economy. This may explain why the District appears to be a community unto itself.

The Eastern Province also shows some interesting differences between the derived eastern communities (communities 4 and 6). The Eastern Province is made up of three districts, which going from north to south are Trincomalee, Batticaloa, and Ampara. Trincomalee seems to have formed its own community (community 4), even extending into a few parts of the North-Central Province and is reflective of more recent economic linkages. What is more interesting and which can probably be answer by historical and economic linkages are what happens to Batticaloa and Ampara. Both of these districts have merged with the Polonnaruwa district in North Central Province to form community 6. This is explained by the fact that the area represented by community 6 is the rice belt of Sri Lanka.

What is represented in Figure 2b are the top-level communities for Sri Lanka. It is possible to zoom-in further by taking each of the identified communities and further decompose them into smaller communities. When we do this for the community formed by Colombo District (i.e. community 7), five sub-communities emerge that are shown in Figure 3.

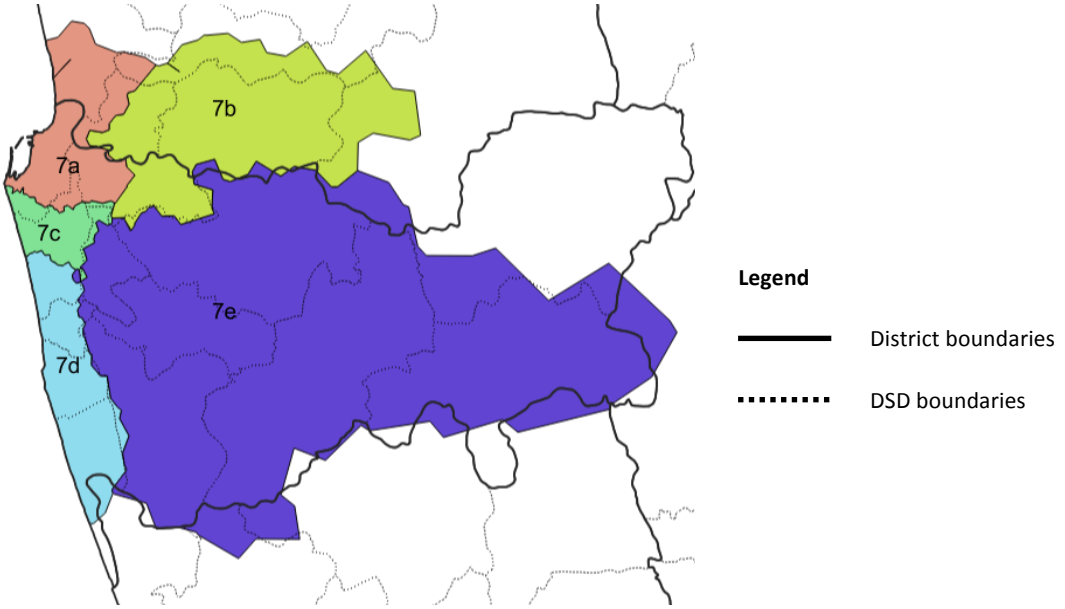


Figure 3: Sub-communities for Community 7 (covering Colombo district)

Even at this level we again find high levels of geographic cohesiveness in the sub-communities. However we find almost no cohesion in the boundaries of these sub-communities with existing administrative boundaries (i.e. DSD level boundaries and even district level boundaries).

5 Conclusion & Policy Implications

Understanding the boundaries of human interactions has implications in a broad range of fields such as social science, disease propagation and economics.

The principle of homophily, the tendency for individuals to seek out and associate with others with similar attributes (eg: location, interests, beliefs, ethnicity etc), is regarded as a major mechanism of social organization. (McPherson, Smith-Lovin, and Cook, 2001) Based on this hypothesis we can consider each of the discovered communities as the result of complex interactions based on different attributes of individuals. The contribution of such different attributes varies from community to community. Previous research has highlighted the role of race and ethnicity in forming social networks in American society (Wimmer, Andreas and Lewis, 2010) The racial and ethnic homophily was found to be strong in a multitude of relationships in the society ranging from marriage (Kalmijn, 1998), to schoolmate friendships (Shrum et al., 1988) to workplace relationships (Ibarra, 1995).

Understanding the spatio-temporal structure of communities as well as the resultant causes for their formation is of particular importance to Sri Lanka, that came out of a three decade long civil war in 2009. The proximate causes of this war can be traced back to ethnic tension amongst the majority Sinhalese from south and minority Tamils in north that started soon

after independence in 1947. The first step in preventing a resurgence of ethnic tensions is to understand the prevalent societal structures and the role of race and ethnicity. Such an understanding is crucial in designing policy solutions to solve this complex problem. With the enforcement of integration laws in public housing projects, Singapore were able to convert its previously ethnically segregated society into an integrated society. (Sim, Yu, and Han, 2003). Whether a similarly strong policy focus on integration is beneficial in the Sri Lankan context is another matter. Irrespective, a policy focus in nurturing society harmony is beneficial. The first step would first involve understanding the existing community structures. The work in this paper is a potential starting point. But further work would be to understand the resultant attributes that contribute to the formulation of each community such as economic, religious, ethnic, class, etc.

The extant administrative boundaries (like in most other countries) were the result of history and geography, and have not changed in more than a century. The only exception is that between 1987 and 2007 the Northern and Eastern Provinces were merged into a single province, and subsequently separated again in 2007. Our work clearly shows that these are in fact separate communities. In fact these two provinces are three communities (community 1,3 and 6 in Figure 2b) or four if we consider the small parts of the Easter provinces that have merged into Community 2. Such findings can have significant socio-political implications for a post-war Sri Lanka. Our findings clearly show that the extant Sri Lankan administrative boundaries do not reflect the natural communities that exist (as inferred by the population's communication patterns). With broader advancements in transport infrastructure (and services) as well as communication, the radii of people's interaction have expanded, defying the traditional borders. Policy makers can potentially utilize such insights to form optimal administrative borders while preserving interaction patterns.

6 References

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J., Rath, C. Ratti, C. (2011). The connected states of America: Quantifying social radii of influence. In *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011* (pp. 223–230). IEEE. doi:10.1109/PASSAT/SocialCom.2011.247

Calabrese, F., Smoreda, Z., Blondel, V. D., & Ratti, C. (2011). Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. *PloS One*, *6*(7), e20814. http://doi.org/10.1371/journal.pone.0020814

Coscia M, Giannotti F, Pedreschi D (2011) A classification for community discovery methods in complex networks. Stat Anal Data Min 4(5):512–546

Fortunato, S., & Castellano, C. (2012). Community structure in graphs. *Computational Complexity*, 490-512.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 1360-1380.

Jung, W., Wang, F., & Stanley, H. E. (2008). Gravity model in the Korean highway. *EPL (Europhysics Letters)*, *81*(4), 48005.

Kalmijn, M. (1998). Intermarriage and homogamy: Causes, patterns, trends. Annual review of sociology, 395-421.

Krings, G., Calabrese, F., Ratti, C., & Blondel, V. D. (2009). Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*(07), L07003.

Ibarra, H. (1995). Race, opportunity, and diversity of social circles in managerial networks. Academy of Management journal, 38(3), 673-703.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415-444.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577-8582.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kerte´sz, J. and Barabási, A.-L. (2007a). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(18), 7332–6. http://doi.org/10.1073/pnas.0610245104

Onnela, J., Saramäki, J., Hyvönen, J., Szabó, G., De Menezes, M. A., Kaski, K., Barabási, A.-L and Kert, János. (2007b). Analysis of a large-scale weighted network of one-to-one human communication. New Journal of Physics, 9(6), 179.

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., … Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PloS One*, *5*(12), e14248. doi:10.1371/journal.pone.0014248

Shrum, W., Cheek Jr, N. H., & MacD, S. (1988). Friendship in school: Gender and racial homophily. Sociology of Education, 227-239.

Sim, L. L., Yu, S. M., & Han, S. S. (2003). Public housing and ethnic integration in Singapore. Habitat International, 27(2), 293-307.

Sobolevsky, S., Szell, M., Campari, R., Couronné, T., Smoreda, Z., & Ratti, C. (2013). Delineating geographical regions with networks of human interactions in an extensive set of countries. *PloS One*, *8*(12), e81707. doi:10.1371/journal.pone.0081707

Thiemann, C., Theis, F., Grady, D., Brune, R., & Brockmann, D. (2010). The structure of borders in a small world. *PloS one*, *5*(11), e15422.

Tinbergen, J. (1962). An analysis of world trade flows. *Shaping the world economy*, 1-117

Wimmer, A., & Lewis, K. (2010). Beyond and below racial homophily: Erg models of a friendship network documented on facebook. American Journal of Sociology, 116(2), 583-642.