Most Calls are Local (but Some are Regional): Dissecting Cellular Communication Patterns

Arvind Narayanan, Saurabh Verma, Zhi-Li Zhang

Department of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota 55414
Email: {*arvind,verma,zhzhang*}*@cs.umn.edu*

Abstract—**We conduct a detailed analysis of cellular communication patterns using (voice/text based) call detail records (CDR) dataset from a nationwide cellular network. We analyze a 5-month large dataset containing over hundreds of millions of CDRs with a user population of over 5 million to dissect meaningful communication patterns, with the goal to understand their impact on – and better manage – cellular network resources. What makes this dataset interesting is that we have both location and timestamp information of the caller and the callee. This allows us to associate communication patterns of users with geographic locations. The enormous size and diversity inherent in the (big)data set, however, makes extracting communication patterns a challenging task. We illustrate this diversity by analyzing tower-level activities and communication patterns between towers and find certain patterns emerging. However, due to the complex structure of the data, extracting them becomes non-trivial. By providing structures to the data in the form of matrices, we adopt machine learning techniques (Laplacian Eigenmaps and t-SNE) to extract "latent" patterns from the data, while accounting for the inherent non-linearity and skewed data distributions. Our main results reveal the existence of interesting regional communication patterns of varying localities and sizes, out of which one pattern scatters across the entire nation. Last but not the least, we also find a number of distinct communication patterns co-existing within the capital city of the nation.**

1. Introduction

Wide adoption of smart phones and other mobile devices has led to rapid growth in mobile traffic, which places a huge demand on the cellular network infrastructure such as resources on cell towers[1], radio network controllers, and so forth. Gaining a deeper understanding of cellular usage patterns and how they are affected by user behavior and mobility is critical to effective management of cellular network resources and to meet user quality of experience expectation. Originally designed for billing purpose, the Call Detail Records (CDRs) collected by cellular network operators provide a useful and rich data source for obtaining insights into network usage patterns and user behavior. Since CDRs are collected at either an initiating cell tower or a terminating cell tower or both (when both caller and callee belong to the same cellular service provider), they allow for more detailed studies of cellular network usage patterns at the (finer-grained) *cell tower* level[2]. In addition, since CDRs are typically stored by cellular network operators for longer periods of time (e.g., for billing), one can also conduct studies of cellular usage patterns over a longer time horizon resulting in the generation of big datasets.

In this paper we study the cellular communication patterns at the *cell tower* level using the CDRs collected by a national cellular provider over a period of close to five months. In particular, using voice calls and SMS messages (thereafter we refer to both simply as "calls") – originating and terminating at cell towers within the same nation-wide cellular provider, we construct the nation-wide (and time varying) traffic matrices at the cell tower level. From a *network-wide* perspective, we leverage the (tower-level) traffic matrices to analyze the usage patterns and geographical distributions of calls at individual cell towers as well as across origin-destination cell tower pairs. We find that call volumes at cell towers are highly diverse, and vary drastically from towers to towers with strong geographical effects. In many towers, a large portion of calls originate and terminate at the same towers, indicating both callers and callees reside within a local area covered by a single tower. On the other hand, the geographical coverage and density distributions of cell towers are highly skewed, with many more towers in large urban areas. While there are strong correlations between geographical distances and call volumes at cell tower level, locality can only explain part of the communication

---

1. In this paper we use the term "cell towers" loosely to refer to radio base stations in a radio access network of a cellular network infrastructure, although we know in reality the radio antenna (typically sitting on a "tower") may not be physically co-located with the actual base station that is attached to, which is the "active" entity that processes user "calls" – including voice, SMS and data.

2. In contrast, since network flow records (e.g., IP/TCP header records captured by Cisco Netflow) can only be collected at SGSN/GGSN (in the case of UMTS 3G networks) or PGW (in the case of 4G LTE networks), without using CDRs or other cell tower-level statistics to map flows to cell towers (which is not an easy task to perform), network flow records only allow for analysis of network usage pattern at the gateway/router level, which is much coarser-grained.

patterns apparent in the cellular call data. To further dissect and extract the significant *latent* communication patterns in the call traffic matrices, we apply the Laplacian Eigenmap method by generating a (high-dimensional) *similarity matrix* from an origin-destination (OD) call matrix based on the empirical distributions of calls from one tower to other towers. This allows us to account for the highly diverse data distributions in the original call traffic matrix, and allows us to extract latent patterns or "clusters" lying in certain lower-dimensional (*non-linear*) manifolds. We also develop a visualization tool to illustrate and interpret the extracted communication patterns.

The main findings of our study are summarized below:

- Although 25% of the total call volumes are generated and consumed by the towers in and around the capital of the nation, the nature of calls, such as incoming versus outgoing volumes *vary* markedly across towers and regions. Not surprisingly, there are *strong dependencies* between call volumes at tower levels and *human activities* around these towers.

- A general observation is that for most of the towers, a majority of the calls are *local*: namely, the caller and the callee for most calls are associated with the same tower, or a *conglomerate* of cellular towers that are located geographically within close proximity. However, we show that such *locality effects* are diverse in that the geographical boundaries are not clear-cut and cannot be simply defined based on geographic distance alone.

- Since our empirical analysis suggests the existence of certain patterns emerging out of the communication between towers, we briefly describe an approach to unravel such hidden patterns by using recently developed state-of-the-art machine learning tools. We find that most of the communication patterns are regional with varying localities and sizes. Moreover, we find that even within the *capital city*, there are a number *distinct "regional" communication patterns*, suggesting the presence of vast diversity in social interactions and human behavior. There is also one communication pattern containing towers that are sparsely distributed across the nation, many of which are located along major transportation networks; this pattern likely captures call activities of long-distance travelers.

The remainder of this paper is organized as follows: Section 2 describes our dataset and the terminologies used in this paper. Sections 3 and 4 analyzes the data in different settings to show the diversity in the data and communication patterns. In Section 5, we briefly describe methods to account for this diversity and to extract meaningful patterns that reveal latent communities of interest from the dataset. The paper is concluded in Section 6.

1.1. Related Work

Call detail records (CDRs) have been extensively used in the past to model and infer user mobility patterns, gain insights about human behavior, and to understand network resource usage. In the last few years, both CDRs and Internet traffic data are exploited to characterize usage patterns and network utility in a cellular network. [1] used CDRs to model user behaviors, but the dataset was limited with just a few hundreds of towers spanning across three weeks. Similarly, [2] exploited CDRs of 35 towers with 15 million voice and 26 million short messages for 60 days to capture usage patterns of users for urban planning. [3] used multi-source data including CDRs to explore human mobility. [4] studied mobile HTTP data traffic in a cellular network and used packet level, flow level, and session level metrics to analyze their characteristics. Authors in [5] used hourglass co-clustering to analyze the traffic data for one day and profile the user behaviors in a large 3G cellular service providers in North America. Similarly, [6] studied the usage patterns of mobile data users using heterogeneous data from one of the largest 3G networks in North America for three months.

Our work differs from the previous studies mainly in two aspects. First, we use a *nationwide* CDR dataset with over 500 million call records spanning across five months, which is much more representative. The enormous size of dataset makes it a *"bigdata" challenge* to extract actionable or meaningful knowledge. While most of the CDR datasets only contain either the source or destination cellular tower information of the call, our dataset consists both the end-points of the call, making it superior for gaining insights about communication patterns across locations. Using this dataset, we are also able to understand the distributions of calls associated with individual towers and between tower pairs from a nation-wide perspective. Second, by using pairwise call volumes between towers, we are able to dissect communities (or cluster of towers) that represent different communication patterns that are reflective of human activities and behavior.

2. Dataset Description

The datasets used in this study come from a national cellular service provider in an African nation. The dataset contain call detail records (CDR) representing voice and text (or SMS) exchanges between subscribers (both of which will be referred to as "calls" in this paper). With more than 500 million records captured over a period of 5 months from over 1000 base stations, our goal is to analyze the data and extract meaningful communication patterns. Every record contains the timestamp of the call taking place, the source base station (or cell tower) from where the call originated and the corresponding destination base station of the call. Additionally, we also have the geographic coordinates of all the base stations spanning across the entire country.

Terminologies: We refer to a cellular base station as a *tower*. When Bob (*caller*), connected to tower A, makes a call to Alice (*callee*) who is connected to tower B, tower A is the *origin* tower, whereas tower B is the *destination* tower. In other words, this call will be considered as an *outgoing* call for tower A, and an *incoming* call for tower
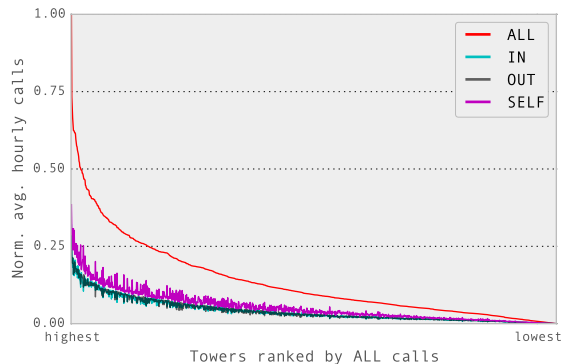
Figure 1: Distribution of `ALL`, `IN`, `OUT`, and `SELF` calls, with a *fixed* order of towers (x-axis) ranked by the `ALL` calls (i.e. total number of calls).

Figure 2: `IN`, `OUT`, and `SELF` call ratio distributions with a *fixed* order of towers (x-axis) ranked by the `ALL` calls (same as in Figure 1).

B. However, if both Alice and Bob are connected to the same tower C, i.e., both the origin and destination towers are the same, then we refer to such a call as a `SELF` call (from the perspective of the tower). Consequently, we use *call direction* to define four aggregated metrics associated with every tower *i*, 1) **SELF calls:** the total number of `SELF` calls for tower *i*, 2) **IN calls:** the total number of incoming calls received by tower *i* excluding `SELF` calls, 3) **OUT calls:** the total number of outgoing calls made by tower *i* excluding `SELF` calls, and, 4) **ALL calls:** the total number of calls seen at tower *i* (`IN` + `OUT` + `SELF`). In this paper the granularity associated with these metrics is *average number of calls per hour*. This was a reasonable choice as we empirically saw (not shown in this paper) that the tower-level average number of calls per hour was highly *correlated* to the total number/volume of calls (i.e. overall tower-level activity).

3. Diversity in Call Volumes

In this section we analyze call volume distributions at tower-level. We try to investigate *"if"* and *"how"* the call volume patterns change over space (i.e. towers) by comparing different types of calls (i.e. `ALL`, `SELF`, `IN`, and `OUT`) with each other.

Call patterns from cellular towers are driven by user demands and behavior. For example, due to its larger population size, we would expect that *as a whole*, towers in urban cities would have higher call volume than the towers located in rural areas. In our dataset, we observe that the capital city of this nation captures more than 25% of the entire call volume. In Figure 1, we rank the towers based on the average number of `ALL` calls made per hour, and plot their distributions based on the call directions. We observe that the distributions of `IN`, `OUT`, and `SELF` calls follow a similar pattern as that of `ALL` calls. In other words, for any tower, the volume of incoming, outgoing and `SELF` calls are correlated to the overall tower-level activity (i.e. `ALL` calls). We further observe that `IN` and `OUT` call volumes are more similar to each other than `SELF` calls. All in all, call volumes
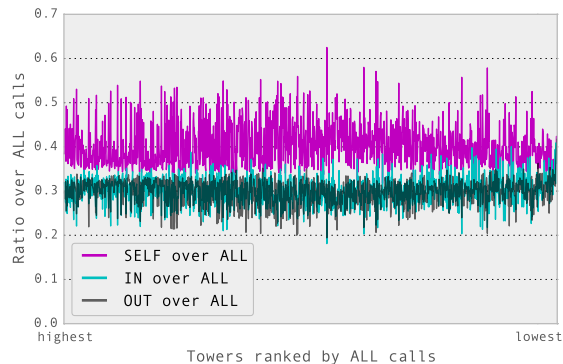
or tower-level activities in general *vary* significantly among different towers.

We now investigate the proportions of `SELF`, `IN` and `OUT` calls over `ALL` calls at the towers. In Figure 2, we fix the rank of towers the same as in Figure 1 and plot the distributions of call proportions – `SELF` over `ALL` (% of `SELF` calls), `IN` over `ALL` (% of incoming calls), and `OUT` over `ALL` (% of outgoing calls). We observe that in general `SELF` over `ALL` call ratios dominate compared to `IN` over `ALL` and `OUT` over `ALL` call ratios, implying people tend to make more `SELF` calls than `IN` or `OUT` calls. In other words, a general trend observed across all towers is that majority of the calls originate and terminate at the same tower. This makes sense as most of the social connections tend to be local to a particular region. However, Figures 1 and 2 show no clear linear relationship between call volume distributions and call proportion distributions. In other words, towers that generate large volume of calls do not necessarily make a lot of calls to themselves. To further investigate, we fix the rank of the towers based on `SELF` over `ALL` call ratio (decreasing order), and plot all the call ratio distributions (see Figure 3). We observe there is still high variance in the call proportions. For example, the `SELF` over `ALL` call proportions vary between 30% to 55%. This implies certain towers tend to make more `SELF` calls than others.

This analysis suggests that there is high variation in the tower-level activities between towers such as the call volumes, where the towers in one capital city consumes more than 25% of such activities. Moreover, we also observe strong *dependencies* between call volumes at tower levels and human activities of either "local" or "mobile" users around these towers. The nature of call volumes (i.e. incoming/outgoing/self) also vary across towers. This is evident by observing the diversity in call proportion distributions among different towers.

4. Diversity in Communication Patterns

From our previous analysis, although we find that majority of the calls originate and terminate at the same tower, we
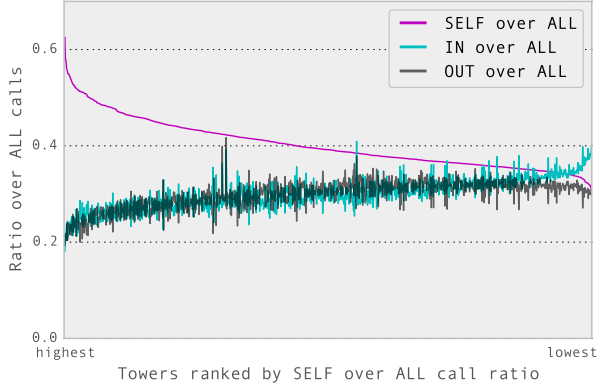
Figure 3: Plot similar to Figure 2, only difference being, the x-axis is now ranked by SELF over ALL ratio (i.e. percentage of calls where both caller and callee both associated with same tower).

could not obtain insights of the *communication patterns*[3] between *different* towers. In this section, we focus on those calls where the origin and destination towers are not the same, and try to find if there are any geo-spatial factors driving communication patterns of such non-SELF calls.
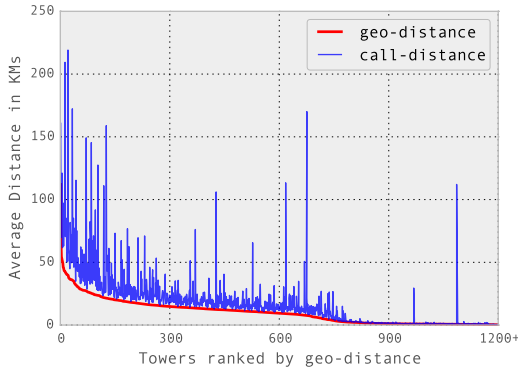

Figure 4: Average distance of each tower with its (geo-distance) and (call-distance) based k nearest neighbors.

4.1. Diversity in the Locality Effects

For every tower, we identify two sets of k-nearest neighbors (kNN) using two different approaches: 1) using geo-spatial distance, and 2) using number of calls, and refer to these sets as G and C, respectively. To elaborate further, G_i is a set of k nearest (or neighboring) towers that are geographically closest to tower i; whereas C_i is a set of k destination towers that tower i makes the most numbers of calls to. For each of these two sets, we then calculate the average geographic

---

3. In the context of this paper, we refer to communication patterns as the tendency to call a particular subset of towers with a higher probability than the rest of the towers. For example, tower A and tower B are said to have a similar communication pattern if both the towers make calls to some subset of destination towers with a similar probability.

---

distance to tower i to compute geo-distance (gd) and call-distance (cd). In other words, $gd_i = \sum_{j \in G_i} dist(i,j) \big/ |G_i|$, where the $dist(i,j)$ is the distance in kilometers (KM) between towers i and j. Since the geographic coordinates *(i.e. latitude, longitude)* of the towers are known *a priori*, we use the *Haversine* formula [7] to compute the required geographic distance between these two towers. Similarly, we also compute $cd_i = \sum_{j \in C_i} dist(i,j) \big/ |C_i|$. In this paper, value of k is set to be 5. We compute both gd and cd for all the towers, and show the results in Figure 4. We fix the rank of towers (i.e. across x-axis) as obtained by the geo-distance gd for all towers, largest to smallest. We can clearly see that overall there is slight level of correlation between both gd and cd for certain set of towers. In other words, certain localities (or towers) tend to make more calls to towers that are regionally closer to them, there by showing some *"locality"* effect. However this is not applicable across all the towers, and this *diversity* in *locality effects* can be seen as *fluctuations* in the plot corresponding to call-distance in Figure 4. While this maybe a side effect to choosing k=5, our objective was to show the diversity in these relations. However this gives us an intuition of the existence of certain communities of people (i.e. collection of towers) that tend to talk with each other more than others. In Section 5, we briefly describe an approach to identify such communities, and also show the results obtained from this dataset.
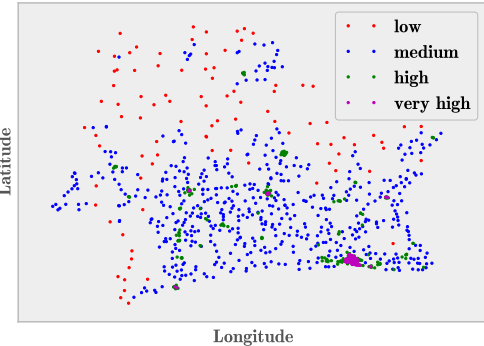

Figure 5: Geographical map of towers classified using geo-spatial neighborhood tower-densities (or *geo-distance*).

4.2. Diversity in Spatial Tower-Densities

By focusing only on the geo-distance plot (i.e. red line) in Figure 4, we make another observation that relates to the diversity in spatial (or neighborhood) tower-densities. The first few towers ($\sim$1 to 150) have very high geo-distance, which indicates such regions have very low regional tower-density. Going beyond the 150th tower, we see a steady yet slowly decreasing geo-distance from towers 151 to 650. Then there is a drop from $\sim$650-to-800th tower, and finally the geo-distance is at its lowest beyond the 800th tower. This last segment corresponds to regions where the tower density

is at its highest, which most likely represents the urban (or metropolitan) areas of the country. To validate this, we roughly come up with 4 contiguous ranges of geo-distance to classify the towers into 4 categories based on their geo-spatial tower-density – *low*, *medium*, *high*, *very high*. We then map all the towers on a geographical map labeled with their category in Figure 5. Further investigation of the geographic properties and the obtained clusters verifies our earlier observations that the towers with *very high* tower density represent the biggest (or the most populated) urban cities. Towers labeled as *high* (i.e. green dots) represent the suburban areas surrounding the bigger cities, as well as some of the tier-2 cities. Towers with blue dots (labeled as *medium*) represent the rural and transit areas. Finally red towers represent the remotest parts of the nation. In a nutshell, this analysis suggests that the diversity in tower-density highly depends on geographic-specific properties such as human population, user demands. However, it is not clear about the effect of such varying levels of tower-density on the call volume (or tower-level activities). In the next subsection, we dwell in this particular direction.
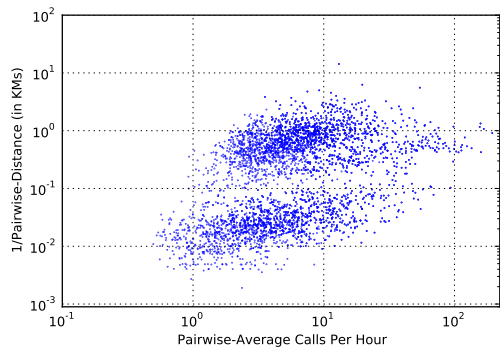

Figure 6: *log-log* scatter plot showing the relation between a tower's location and its average in-out calls.

4.3. Spatial Diversity in Pairwise-Call Volumes

Next, we consider volume of calls between towers, i.e. *pairwise-calls*, and investigate its relation with the distance between the paired towers. For this part of the analysis, we pair every tower with 5 other towers that they mostly interact with. We use the same *Haversine* formula [7] to compute distances between towers. Figure 6 shows the scatter plot (*log-log* scale) of this relation. Note, the y-axis is the inverse of the pairwise-distance. We observe certain pairs of towers that are geographically close to each other generate large call volumes. At the same time, there are some other pairs of towers that are located relatively less closer also seem to generate large call volumes. Nonetheless, we do see certain patterns emerging from the scatter plot. However, high number of towers coupled with the observed spatial diversity in pairwise-call volumes makes it difficult to extrapolate the relations.

In this section, we observe that there exist locality effects in communication patterns among towers along with some

diversity. Our analysis also show that the tower density across the nation has a close relationship regional geographic features such as whether it is a metropolitan area or a transit region. Last but not the least, irrespective of two towers being geographically close or relatively far, for the both cases we find them to have the ability to generate large call volumes.
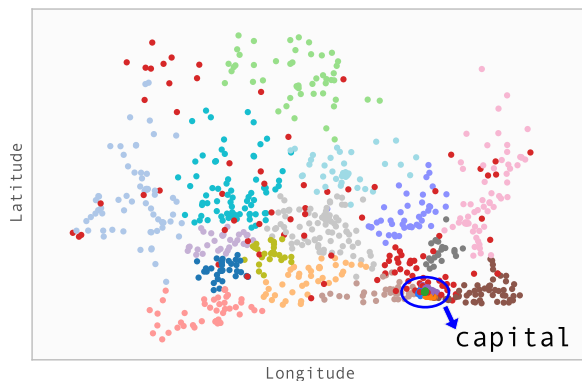
5. Extracting Regional Communities

So far, we empirically analyzed the dataset to show that although most of the calls are local to towers, we still observe groups of nearby (or neighboring) towers communicate with each other more than the rest of the towers. In other words, there are communities (or collection of towers) that tends to communicate more among themselves and at the same time show similar behavior to communicate with others. However, the size or structure of such communities vary. For example, earlier we tried using kNN, however fixing the value of k to be a constant does not work well due to the diversity in data. Although not shown, we find that linear-dimension reduction techniques such as PCA are ill-suited for this dataset as the number of components with significant eigenvalues is very high (greater than 80) suggesting the data is high dimensional and diverse (thus, *non-linear*) in nature. Hence, there is a need to account for this diversity and come up with an approach to identify communities (or clusters) of towers having similar communication patterns.

Dest. Origin	tower A	tower B
tower A	SELF calls	A to B calls
tower B	B to A calls	SELF calls

Table 1: Origin-Destination (OD) Matrix Representation

We combine some of the popular algorithms with state-of-the-art machine learning tools to develop an approach to account for such non-linearity in data and extract clusters (or "latent" patterns) arising from such datasets. The key idea is the following: Instead of directly working on the original dataset, we construct an origin-destination (OD) matrix such that each row represents a data point (say, the origin tower) and treat every column as a feature (or destination). Therefore, a matrix of size $n \times m$ represents n data points, where each data point is a vector of m features (see Table 1). Each cell in the matrix corresponds to the *number of calls* from origin an tower (i.e. row index) to some destination (i.e. column index). The diagonal elements denote `SELF` calls. This matrix represents an *empirical* distribution to account for the highly diverse data distributions in the OD matrix. From such distributions, we then derive a new (now symmetric) *similarity* matrix of size $n \times n$. This amounts to transforming the data points in original OD matrix into a kernel space via a (non-linear) Gaussian kernel function. We apply the Laplacian Eigenmap method to cluster data points and extract patterns or "clusters" lying in certain lower-dimensional (non-linear) sub-manifolds. To

(a) Cluster results overlaid in a geographic map of the nation.
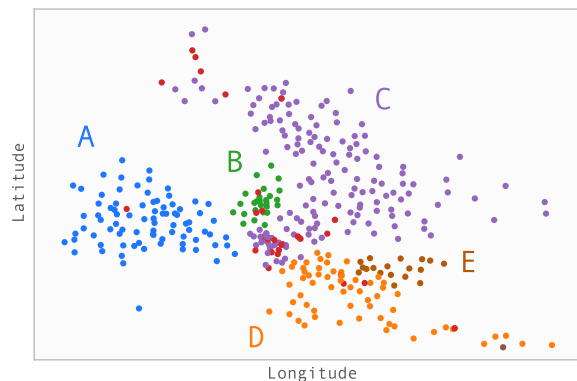
(b) Cluster results overlaid in a geographic map of the capital city and its close vicinities.

Figure 7: Clustering results of towers show formation of *regional* communities in communication pattern across the nation. However, one of the pattern (red colored towers) travel across the entire nation. (*best viewed in color*)

demonstrate the inherent non-linearity of these latent clusters and to help interpret the results, we visualize the data points by projecting them into a 2-dimensional space using t-SNE method [8]. We leave a more detailed analysis and theoretical-based discussion about this approach as part of a longer version of this paper.

Figure 7a shows the result of applying this approach on the CDR dataset. We found a total of 21 clusters, where each cluster contained a certain subset of towers. Note, our approach will put two towers in the same cluster if both the towers have a similar call-distribution to destination towers. Therefore, the extracted clusters represent 21 distinct *communication patterns* in the dataset. All the figures are best viewed *in color* – each color represents a cluster. Except for the cluster denoted using red color (●), all other clusters represent *regional* communication patterns of varying localities and sizes. It is interesting to see such results emerge, especially when no geographic information was explicitly mentioned in the input to our pattern extraction methodology. These regional communication patterns capture social interactions and human mobility in this African nation. As clearly shown in Figure 7a, users tend to interact (in this case, call or message) with others in certain geographical regions. In fact, further in-depth analysis suggests vast swath of the nation (outside the capital city) can be divided into a few distinct communication "zones" where users tend to interact more with others in the same zone, or when interacting with users outside their zones, they tend to have similar communication patterns. We zoom into the results of the capital city (see Figure 7b). It is interesting to observe that the capital city is itself dominated by five distinct communication "zones": cluster C (●) which is the largest, cluster B (●), cluster A (●), cluster E (●) and cluster D (●) which are very close to each other. Further investigation of the geographical features of these clusters reveal that in fact the capital city consists of a large mainland area (where most towers in clusters A, B and C reside) and two (connected) islands separated from the mainland by a bay. Clusters D

and E have distinct communication patterns from cluster A, B and C, as residents in the islands also tend to work on the islands; hence most calls are confined within the islands. In contrast, clusters A, B and C represents call patterns not only among residents within the capital city, but also that these residents tend to interact with users residing in many other surrounding areas. Finally, the towers in red-colored cluster (●) are sparsely distributed across the nation, most of which have relatively low overall call volumes and many are located along major transportation networks. This suggests towers in the red cluster captures call activities of users in transit across the nation.

6. Conclusion

In this paper, we presented our detailed analysis on a nationwide large CDR dataset enriched with both source and destination cellular tower information of over 500 million calls and found some interesting patterns. We analytically observed that there is a wide diversity in call volume patterns and there exist a strong dependencies between call volumes at tower levels and local (or mobile) users around these towers evident from call proposition distribution among different towers. We also found the diversity in communication patterns in terms of locality and regional effects along with variation in tower density dominated by urban areas across the nation. Unraveling patterns from such an enormous dataset coupled with diversity constitutes a bigdata challenge. To address this problem, we provide structure to the data by constructing origin-destination (OD) matrix and employed advance machine learning tools (Laplacian Eigenmaps and t-SNE) to find, and also importantly understand, the hidden patterns from the data. Our main results reveal the existence of regional communication patterns of varying localities and size in cellular network, out of which one spanned across the entire nation. As a part of future work, we plan to make our approach more general by showing its efficacy in different application domains by accounting

datasets which do not semantically fit within our notion of origin-destination (OD) matrix structure.

Acknowledgments

References

[1] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," in *New frontiers in dynamic spectrum access networks, 2008. DySPAN 2008. 3rd IEEE symposium on*.   IEEE, 2008, pp. 1–11.

[2] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 0018–26, 2011.

[3] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proceedings of the 20th annual international conference on Mobile computing and networking*.   ACM, 2014, pp. 201–212.

[4] Y. Zhang and A. Årvidsson, "Understanding the characteristics of cellular data traffic," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 461–466, 2012.

[5] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3g network using hourglass co-clustering," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*.   ACM, 2010, pp. 341–352.

[6] Y. Jin, N. Duffield, A. Gerber, P. Haffner, W.-L. Hsu, G. Jacobson, S. Sen, S. Venkataraman, and Z.-L. Zhang, "Characterizing data usage patterns in a large cellular network," in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*.   ACM, 2012, pp. 7–12.

[7] C. Robusto, "The cosine-haversine formula," *The American Mathematical Monthly*, vol. 64, no. 1, pp. 38–40, 1957.

[8] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research (JMLR)*, 2008.