Communication flow patterns in the D4D dataset

Paul Schmitt, Morgan Vigil, Mariya Zheleva, and Elizabeth M. Belding
Department of Computer Science University of California, Santa Barbara
{pschmitt, mvigil, mariya, ebelding}@cs.ucsb.edu

Abstract

Traffic analysis of mobile and Internet networks helps researchers understand people's behavior and needs in the context of these networks. Such analysis is an important facet of both the initial design as well as the iterative improvement of applications that leverage such networks. In developing countries where the population is predominantly rural, mobile communications with their high affordability and intuitive interface, are the first communication technology introduced. Thus, the analysis of usage patterns of mobile networks is of great importance, as it facilitates better understanding of people's interaction with technology and their specific technological needs. We approach the D4D challenge as a preliminary analysis on network usage patterns focusing particularly on usage in Rural areas. We also analyze persistence trends of individual's social groups in this mobile network. Based on our results, we provide a discussion of possible practical applications that can leverage mobile networks.

1 Introduction

Mobile networks have revolutionized the way people communicate in the developing world and serve as a platform for enhancement of many aspects of people's day-to-day life. Applications that use underlying mobile networks span from health care [5, 6, 11, 15] and education [18, 1, 14] to agriculture [7, 17, 16] and mobile banking [12]. Multiple successful projects in Africa have spurred from observing people's behavior in mobile or social networks. Johnson et. al., after analysing facebook traffic, design a system to facilitate local content sharing within remote rural communities [9]. [12] describes a system called mPesa 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 indicated 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 people no longer needed to carry cash.

Such projects are of critical importance to introducing new services and enhancing the wellbeing 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.

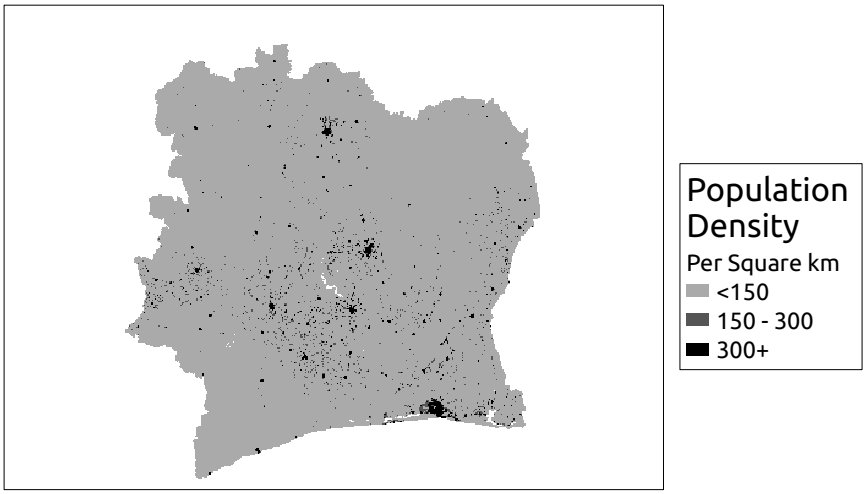
Due to the prevalence of mobile communication technologies in sub-Saharan Africa [13], it is particularly important to understand how information is exchanged over these wireless infrastructures. As communication patterns emerge, it becomes possible to improve current wireless infrastructures and develop new technologies and systems that effectively leverage existing infrastructure. We are particularly concerned with how communication patterns might inform healthcare systems for development. We approach Orange's datasets on mobile call patterns in Côte D'Ivoire with this end in mind.

Our previous work introduces VillageCell [4], a comprehensive connectivity solution designed for rural areas. VillageCell includes low cost free-to-air cellular coverage and exploits locality of interest to facilitate effective use of limited gateway bandwidth. VillageCell has been deployed in rural Macha, Zambia. By understanding how mobile communication flows with respect to population density, we begin to understand the feasibility of a similar connectivity solution for rural Côte D'Ivoire. Additionally, we investigate how mobile communication flows socially. Social graph patterns that emerge provide a preliminary understanding of how social relays might be exploited for information broadcast. We discuss how this can be useful to the development of a healthcare system that uses cellular network technology.

2 Methodology

To facilitate extensive analysis of mobile network traces over different subsets of the Côte D'Ivoire population, we preprocess our data. We now describe the preprocessing techniques and software we use. We also define a set of metrics we utilize in the process of evaluation.

Figure 1: Population density of Côte D'Ivoire.



2.1 Datasets

Our analysis focuses primarily on information in Set 1 and Set 4 of the Orange Datasets. Dataset 1 represents mobility data aggregated on an hourly basis for ten weeks from December 05, 2011 to April 22, 2012 and includes data about the number of calls and call duration between antenna pairs. In addition, we use the antenna location data provided by the ANT_POS dataset. This set provides data that maps an antenna ID to its corresponding latitudinal and longitudinal coordinates. Set 4 includes ego-centric social graphs that describe up to second order neighbors of 5000 users traced over the entire period. We use this dataset to analyze persistence of social groups that a mobile subscriber communicates with.

As recommended by Orange, we use data from the AfriPop project. The data consist of high resolution population density distribution information in ESRI Float format. We use this data to calculate and associate population density with antenna locations given the ANT_POS dataset.

2.2 Antenna classification

We employ the new European Union typology of "predominantly Rural", "Intermediate", and "predominantly Urban" areas. This typology is a modification of the Organisation for Economic Co-operation and Development (OECD) methodology that seeks to minimize distortions caused by large variations in the area of local administrative units [2]. Using the new OECD method, rural local administrative units are defined as areas with a population density below 150 inhabitants per km^2 applied to grid cells of size 1 km^2. Likewise, urban local administrative units are defined as areas with population density of at least 300 inhabitants per km^2 applied to grid cells of size 1 km^2. The 1 km^2 cell size provides fine granularity, which makes the OECD method equally applicable to countries outside the European Union. See Table 1 for the

Table 1: OECD population density classifications

Density per km^2	Classification
0-149	Rural
150-299	Suburban
300+	Urban

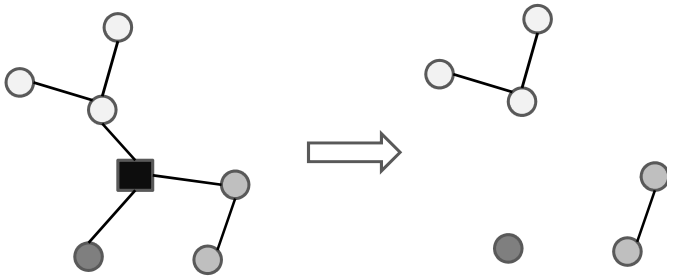


Figure 2: The effect of removing the ego (depicted with a square) from the ego-centric social graph.

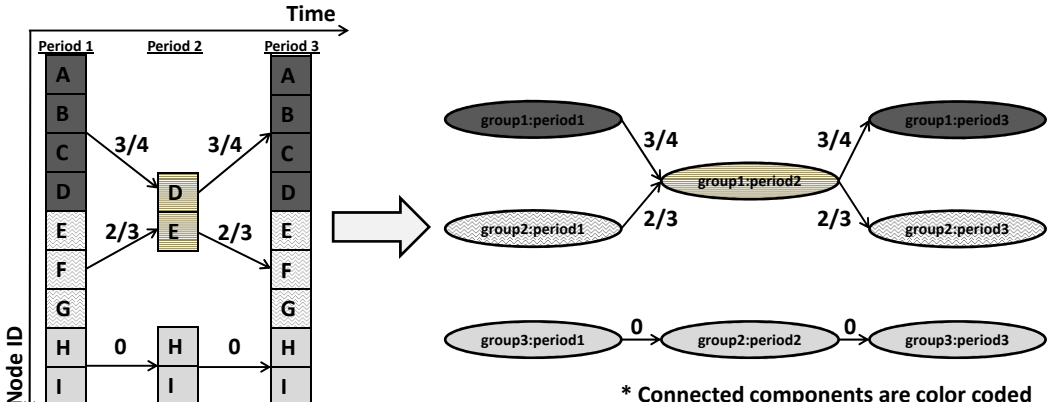


Figure 3: Building a persistence graph.

classifications per km^2 we employ. Note that our classification of "Suburban" directly corresponds to [2]'s classification of "Intermediate".

We utilize the population density information contained in the AfriPop data set and use Quantum GIS to project it as a raster layer. The AfriPop data includes population density information formatted as the number of people per 100 square meters for Côte D'Ivoire. We then re-sample the density data at a lower resolution creating a grid of 2 km squares assigning population density for each as the mean density values of the AfriPop data bounded by the new grid. Each square is assigned to one of the population density categories using the OECD typology. This allows us to create the population density map shown in Figure 1. The different shades indicate the different density classifications as defined by OECD - Rural, Suburban, and Urban. As shown, the majority of land area in Côte D'Ivoire is classified as Rural. Interestingly, World Bank [3] population statistics show that the rural population represents 51.2 percent of the total population in the country while the urban population accounts for 48.8 percent. Our grid assigns population density at a resolution suitable for associating antennas with the underlying population statistics.

2.3 Ego-centric graphs analysis

We examine the ego-centric social graphs dataset to determine persistence of social groups for each ego over time. We also analyze the likelihood that one or few nodes – *top nodes*, persist over time in an ego-centric graph. We hypothesize that such persistent nodes can be used as information relays in an egocentric graph. Our analysis indicates that such persistent nodes indeed exist. Their feasibility as information relays, however, needs further analysis that requires richer datasets providing information about frequency and duration of phone calls as well as physical location of the communicating parties.

$$J = \frac{A \cap B}{A \cup B} \tag{1}$$

The Jaccard similarity changes between 0 and 1, where 0 indicates no overlap and 1 indicates full overlap.

In order to extract the separate social groups of an ego, we remove the ego node from each ego-centric social graph

(Figure 2) and analyze the connected components that remain after the ego is removed. Each connected component corresponds to one social group. Note that in the text we use the terms *connected component* and *social group* interchangeably.

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 the same 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. J for two sets A and B can be calculated as follows:

Figure 3 presents an example of building the persistence graph for a single ego over three consecutive periods. The left-hand 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-hand side of the picture presents the resulting persistence graph. 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 "group1:period1" and node "group1:period3").

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.

Further 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 $\hat{JS}^{(t,t+1)}$ 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.

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

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.

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

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

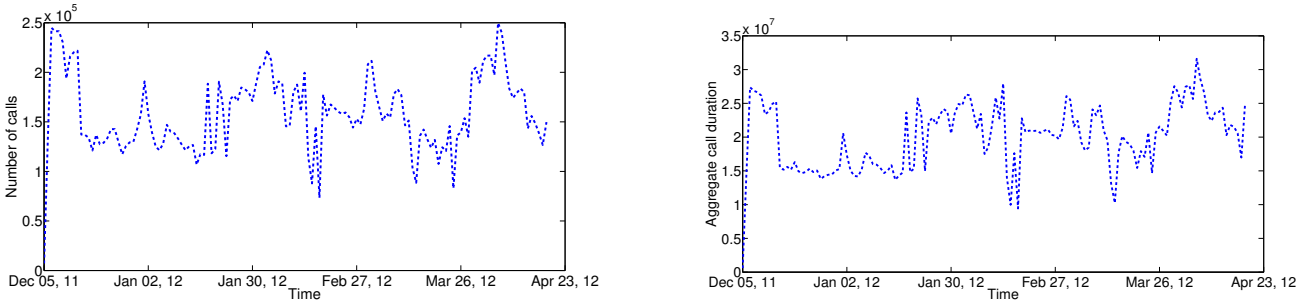
We present our results for social groups persistence in Section 3.6.

3 Evaluation

We begin our analysis by investigating temporal trends in mobile communication in general and across areas of different population density types. We expect to observe regular temporal trends along weekly and monthly intervals, with Rural areas having temporal trends distinctive from those of Urban areas. Progressing from temporal trends, we explore trends related to population density. Again, we expect to see differences in call duration and call frequency based on the population density of the sender. Next, we seek observable relationships between the distance between sending and receiving antennas and call duration and frequency. Finally, we examine patterns in social groups. We hope to observe consistency in social groups over time.

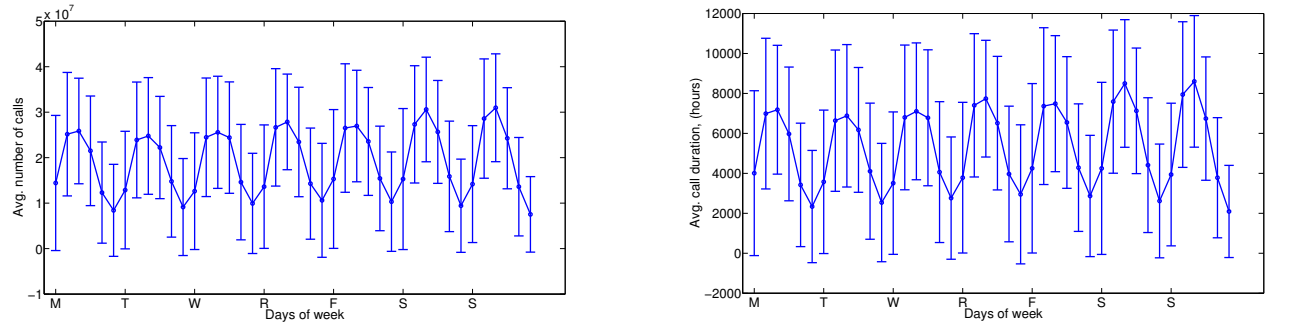
3.1 Usage patterns over time

We evaluate the cellular network activity patterns over the entire capture period. In Figures 4(a) and 4(b) we plot aggregate number of calls and call duration per day. As the figure shows, there is no distinctive call pattern on a weekly or monthly basis; instead subscriber activity seems to be widely correlated with events in the country. We hypothesize that the peaks from the beginning of the period coincide with the weeks before and after the parliamentary elections on December 11th, 2011, while the second peak is most likely traffic around the New Years Eve. The increased utilization



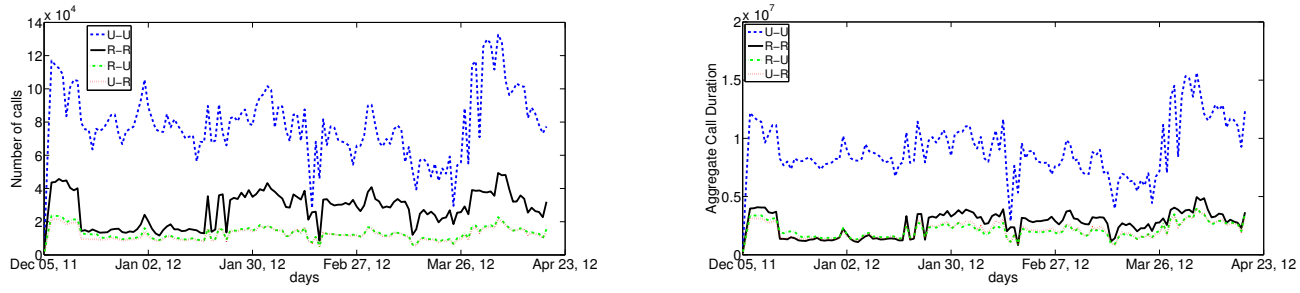
(a) (b)

Figure 4: (a) Number of calls and (b) call duration over time.



(a) (b)

Figure 5: (a) Number of calls and (b) call duration over time.



(a) (b)

Figure 6: (a) Number of calls and (b) call duration over time.

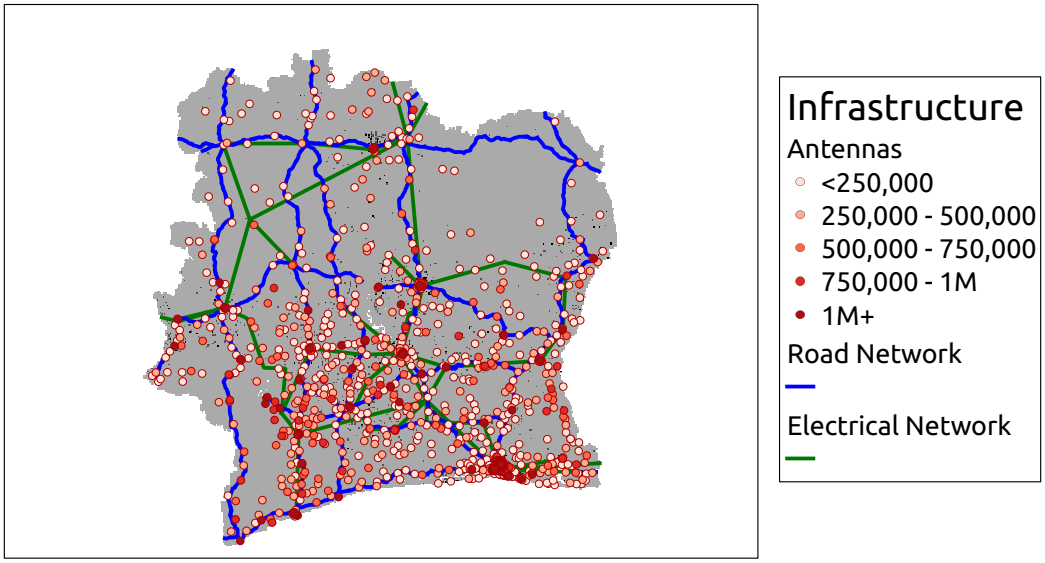


Figure 7: Antenna Usage

from the end of March and April is likely associated with the military coup in Mali and the associated ECOWAS[1] summit that took place in Abijan, Côte D'Ivoire. Such irregular usage pattern is very different than what had been observed in cellular network traces from the western world [10].

The lack of weekly pattern is further confirmed by Figure 5. We average the number of calls and call duration over the entire capture period in a one week window. Each point on the plot presents an average over four hours over all occurrences of each day of the week (that is the first data point from the graphs presents the average number of calls and call duration for the hours from Midnight to 4 AM for all Mondays in the capture period). The figure clearly presents diurnal pattern of network activity with slight increase over the weekend, however, the standard deviation of this graph is very high, indicating that the network activity varies dramatically over the observed period.

Next we analyze whether the calling patterns of rural areas differ from those in urban areas within Côte D'Ivoire. In Figure 6(a) and 6(b) we plot the aggregate number of calls and call duration for each day of the observed period. We analyze four categories of calls depending on the source and destination antenna type: Urban to Urban (U-U), Rural to Rural (R-R), Urban to Rural (U-R) and Rural to Urban (R-U). As the figures show, calling patterns for all four categories follow similar trends, where the number of calls and the aggregate call duration between Urban antennas is about three times higher than between Rural antennas. We also note that while the number of Rural to Rural calls is larger than the number of Rural to Urban and Urban to Rural, the aggregate call duration for these three categories is the same. This result indicates that while calls between Rural residents occur more often, they are shorter in comparison to calls between Urban and Rural residents.

3.2 Antenna activity map

We seek patterns of mobile communication flow in Côte D'Ivoire by associating antennas with their geographical location and the population density of their location. The resultant mapping of antennas to location can be seen in Figure 7. We associate each antenna with the underlying population density of their location in order to assign the Rural, Suburban, or Urban typology. Antennas are shaded based on the total number of outbound calls they originate throughout the entire sampling period with darker colors signifying busier antennas. It is evident that antennas are densely clustered in urban

[1] `http://www.ecowas.int/`

locations while more sparsely located in predominantly rural regions. We also find that high activity antennas are often located along major transportation corridors.

Table 2: Antenna Density Classifications

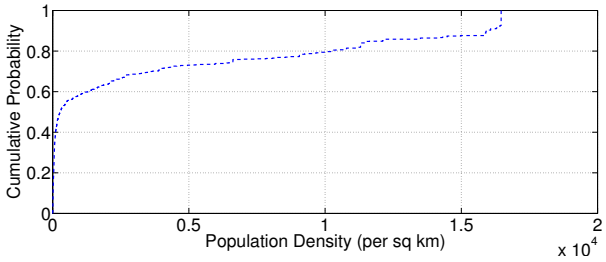
Classification	Antenna Count	Source Calls
Rural	529	146,481,488
Suburban	90	21,529,115
Urban	598	331,630,147
Unknown	21	65,393,926

We join the antenna location and the population density datasets using Quantum GIS and plot all antennas onto a Côte D'Ivoire population density map. This allows us to associate a population density value of the underlying grid with each antenna and assign the OECD typology for each antenna that is provided with geographic information in ANT_POS. Table 2 shows the number of antennas that fall into each of the classifications as well as the total number of calls originated from each antenna type. Of note, we see that relatively few antennas are classified as Suburban. As the antenna location dataset is not fully complete we do not associate any density information for those antennas that do not correlate to a square on the grid. Such antennas are classified as "Unknown" when processing call records.

3.3 Population density

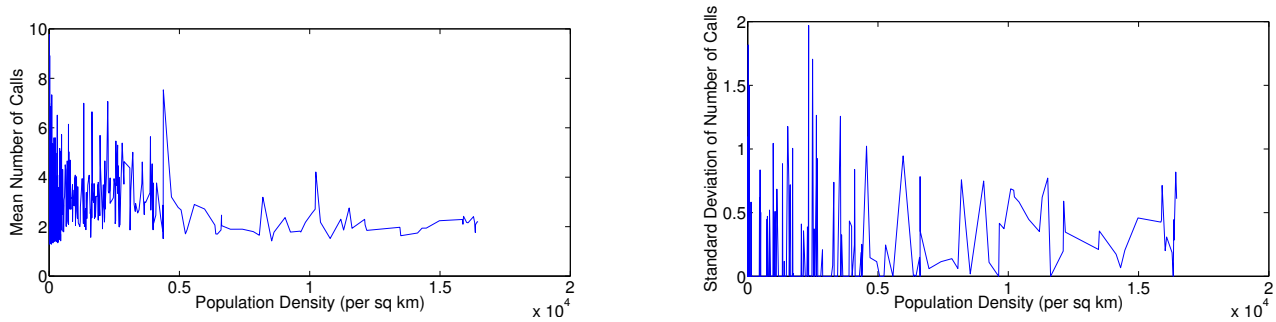
In terms of data density, Figure 8 shows there are observably more records for antenna pairs involving source antennas with population densities with less than 500 inhabitants per km^2. Likewise, Figure 1 shows that the geographical area of Côte D'Ivoire largerly consists of sparsely populated areas regions. This leads us to examine the distribution of Set 1 in terms of population density. Additionally, the distribution shown in Figure 8 demonstrates a clear dichotomy between densely populated regions and sparsely populated regions. This leads us to classify antennas into one of three population density categories: Rural, Suburban, and Urban. In Seciton 3.5, we then classify each identifiable recorded connection in Set 1 based on its directionality: Rural to Rural, Rural to Suburban, Rural to Urban, Suburban to Suburban, Suburban to Rural, Suburban to Urban, Urban to Urban, Urban to Suburban, or Urban to Rural.

Figure 8: Distribution of population density (per sq km) associated with source antenna.



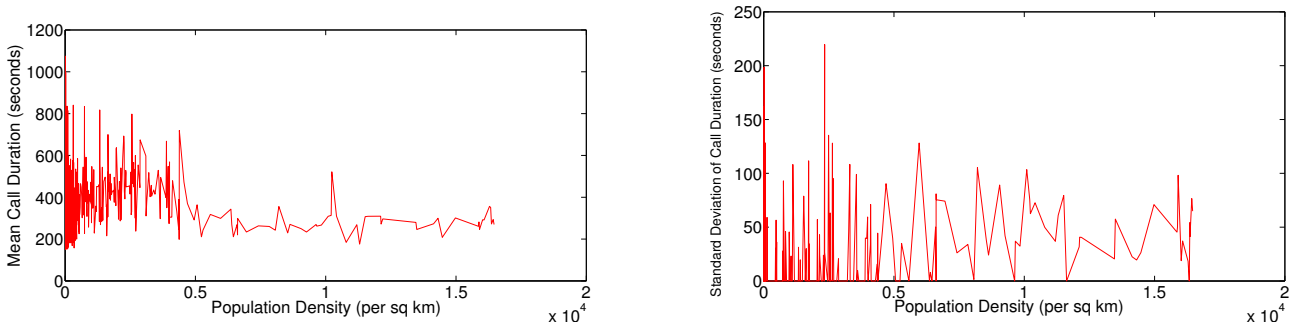
As evidenced by Figure 1, population density in Côte D'Ivoire varies between Rural and Urban areas. We explore the relationship between the population density of a sending antenna and the average number of outbound calls associated with the antenna. Because of the predominant use of "Calling Party Pays" (CPP) policy in sub-Saharan Africa, we focus on the number of outbound calls sent from an antenna rather than incoming calls received by an antenna [13][8]. According to Figure 8, it appears that a large cluster of data points occur in population densities below 500 inhabitants per km^2 and a smaller cluster occurs in population densities above 15,000 inhabitants per km^2. In order to normalize for this, we calculate the mean number of calls and the mean call duration for population densities per km^2 in Figure 9(a) and Figure 10(a). The mean values in Figure 9(a) illustrate the tendency for the outbound number of calls to decrease as population density associated with the sending antenna increases. Likewise, Figure 10(a) illustrates the tendency for the call duration to decrease as population density associated with the sending antenna increases. Due to the CPP policy, we anticipated a larger mean number of outbound calls and mean call duration from antennas associated with the highest population density, which is associated with Côte D'Ivoire's financial district and center of commerce. In order to explore why the mean number of outbound calls and call duration were greater for lower population densities, we plot

the standard deviation of the number of outbound calls and call duration in Figure 9(a) and Figure 10(b). We observed more variation in the standard deviation values associated with lower population densities. We attribute the variation of standard deviation of the number of outbound calls and call duration for lower populaiton densities to be indicative of sparse antenna placement. For instance, many of the lower population densities associated with low standard deviations of number of outbound calls and call duration only have one or zero antennas associated with them. This makes it more difficult to ascertain a "normal" call pattern for areas of low population density, even though most of the geographical area has low population density values (see Figure 1). However, Figure 7 may demonstrate why we observe such erratic mean number of outbound calls and mean call duration in areas of low population density. As can be seen in Figure 7, there are several antennas associated with low population densities that have a high number of outbound calls associated them. In Section 4 we discuss how infrastructure can be inferred based on population density and call frequency associated with antennas.



(a) (b)

Figure 9: (a) Population density vs. mean number of outbound calls and (b) Population density vs. standard deviation of number of outbound calls



(a) (b)

Figure 10: (a) Population density vs. mean call duration and (b) Population density vs. standard deviation of call duration.

3.4 Mean call duration as a function of distance

We investigate the relationship between call distance and the average duration of calls. We calculate the distance in kilometers between all pairs of antennas with known geographic location using the Haversine formula [19] and inputting a mean earth radius of 6,372.80 km. We group connection distances into the nearest 10 km in order to calculate aggregate statistics for each group.

Next we calculate the mean call duration for each of the distance groups. We process Set 1 to find distance information for each record and associate call duration and call counts to the associated distance. Records that include antennas for which we do not have location information are ignored. The impact of distance between source and destination antenna on mean call duration is seen in Figure 11. In general, we observe an increase in average call durations as connection distance increases. We hypothesize that the reason for such a pattern is the calling parties have fewer opportunities for

in-person interactions due to the geographic distance. Lastly, note that with relatively fewer call records for distances greater than 500 km, more noise is introduced into the graph. Given more call records we expect that the relationship trend between distance and average call duration would hold.

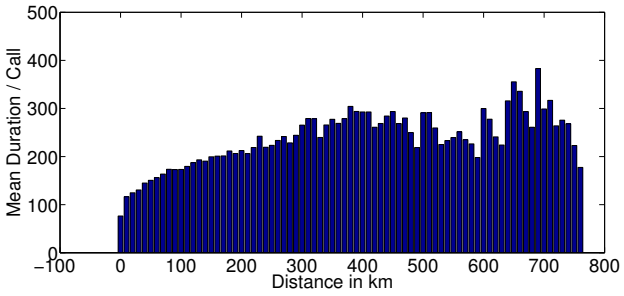


Figure 11: Antenna distance vs. mean call duration.

3.5 Call typology classifications

We investigate the potential correlation between population density and calling patterns by associating antennas with known locations to the corresponding local population density. This process yields antennas denoted as Rural, Suburban, Urban, or Unknown for the antennas which have no geographic location. We process Set 1 to classify call records by each typology source and destination pair in order to investigate potential communication patterns. In this analysis we do not consider records for antennas with no geographic data or records without valid antenna IDs. As seen in Figure 12, the majority of the identifiable connections are classified as Urban to Urban connections. This is followed by 20% of connections classified as Rural to Rural. Connections classified as Rural to Urban or Urban to Rural each account for 9% of connections.

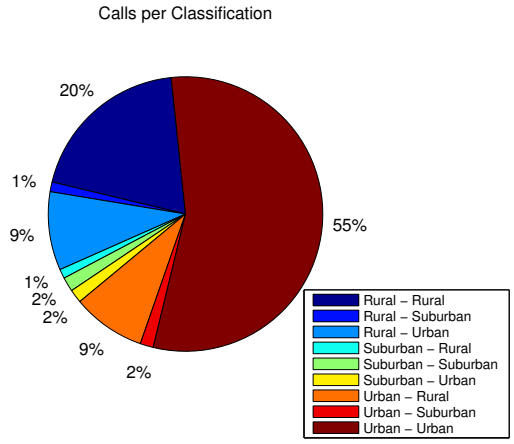


Figure 12: Classification of communication between antenna pairs.

Next, we search for differences in mean call duration across the connection classifications with results shown in Figure 13. We find that the two call classifications with the longest mean call duration are Urban to Rural and Rural to Urban. An observable phenomenon is that calls confined to the same source and destination density type are noticeably shorter on average compared to calls between mixed pairs. Given our prior finding of the relationship between call distance and average duration we posit that the majority of calls that do not cross classification boundaries are confined to a smaller geographic region. For instance, we believe Urban to Urban calls are more likely to be sourced from and destined for the same urban area. Lastly, an interesting observation is that calls originating from Urban antennas generally have a longer duration for any destination type. We believe this is due to the common policy of "Calling Party Pays" and higher buying power of individuals who reside in urban areas.

This trend leads us to look at the average distance between connecting antennas associated with each connection density classification type. Figure 14 shows the relationship between the average call distance for each connection classification type

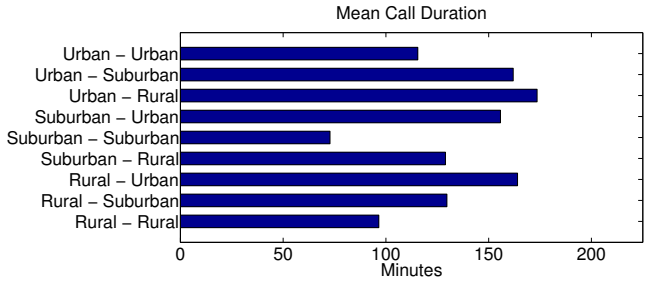


Figure 13: Call durations for classified connections.

and the average call duration. We see that the longest average distance between connecting antennas occurs between Rural to Urban and Urban to Rural connections. The shortest average distance occurs between similar-to-similar connections such as Rural to Rural, Urban to Urban, and Suburban to Suburban. As we would expect based on Figure 11, we see classifications associated with longer average distances between connecting antennas also associated with longer average call duration.

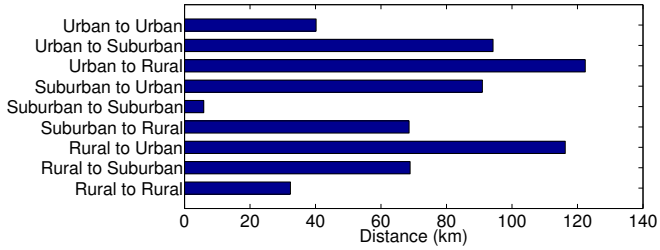


Figure 14: Mean distance for classified connections.

We investigate associating call patterns and population density for calls that are sourced and destined for the same antenna ID. This is motivated by the observation in Figure 12 that roughly 75% of all observed calls are categorized as Rural to Rural or Urban to Urban. We process Set 1 and identify records that include two valid antenna IDs where the source and destination match. We find in Table 3 that 57% of all Rural to Rural calls are both sourced from and destined for the same antenna. We posit that this is due to fewer available antennas to associate with in predominantly rural areas. Furthermore, the cellular coverage provided by a single antenna in rural settings is typically larger, which means that a higher proportion of local users are associated with the same antenna. Calls between users in the same general vicinity in a Rural area are likely to involve only one antenna. Interestingly, Urban connections sourced from and destined for the same antenna represent 23% of all Urban to Urban calls. We believe that the higher density and smaller cell range of Urban antennas provides more diverse antenna association possibilities for users.

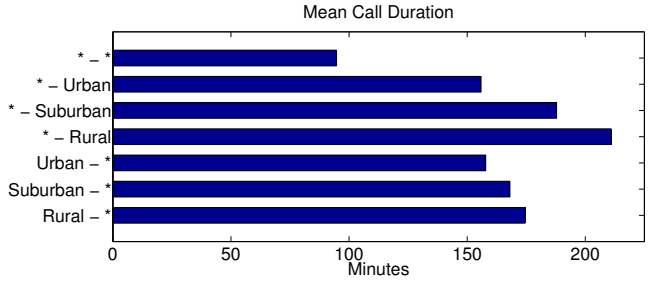
Table 3: Percent of calls made between same source and destination antenna

Classification	Percentage of calls
Rural to Rural	57%
Suburban to Suburban	88%
Urban to Urban	23%

Our final analysis is focused on antennas for which we have no population density information. These antennas include those that are not provided with geographic coordinates in the ANT_POS data set as well as those identified in Set 1 with an invalid ('-1') antenna ID. We classify call records from Set 1 where at least one antenna in the connection is a part of the "Unknown" antenna classification and gather statistics. Figure 15 illustrates that calls between two unclassified antennas are typically shorter than those in which one side of the connection is "known." Given our prior observation that mean call durations are noticeably shorter when the source and destination antenna classification is not mixed we believe

that calls between two unclassified antennas remain within the same density classification, though unknown. Analysis of records in which one of the involved antennas is classified as known shows that calls involving a Rural antenna for at least one half of the connection are longer on average than other types. Also of note is that calls with one half of the connection known are more similar to patterns associated with mixed classification calls than those remaining within the same classification.

Figure 15: Mean call durations for connections including unclassified antennas



3.6 Egocentric graphs

We now examine the ego-centric social graphs provided in dataset 4. Our analysis focuses on persistence of social groups with which individual egos communicate. We regard this analysis as preliminary work on identifying persistent neighbors within one's social network, who can serve as reliable information relays.

First we provide high level analysis 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 (i.e. 10). Figure 16(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, the majority of egos – 68%, have between 2 and 5 connected components on average. Further, we examine how the number of connected components deviates for each ego. Figure 16(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 ego-centric 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 2.3, a node in period t would have 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 would 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 17(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 did not occur in the preceding and following periods. 40% of the nodes have in- or out-degree of 1, which means that these 40% of the social groups persisted in consecutive periods. Only about 3% of the cases have in- or out-degree larger than 1, which means that social groups rarely split or merge over consecutive periods.

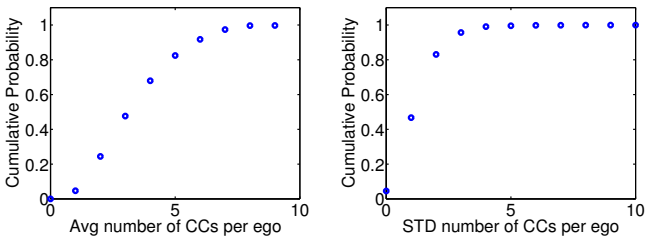
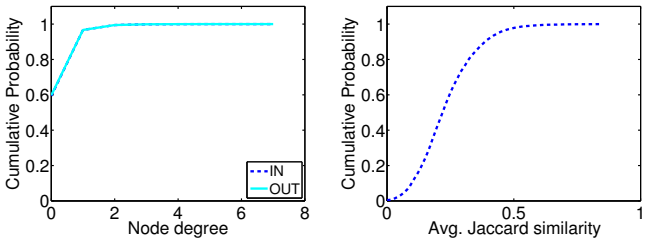


Figure 16: (a) The number of connected components per ego and (b) the standard deviation of the number of connected components per ego over the observed period.



(a) (b)

Figure 17: (a) The in- and out-degree of nodes in all persistence graphs and (b) the average Jaccard similarity for each persistence graph.

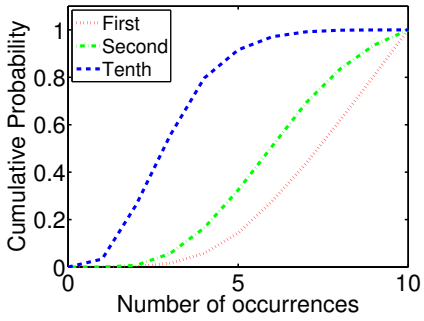


Figure 18: Number of occurrences of the first, the second and tenth most frequent neighbor.

This result indicates an important quality of the observed ego-centric 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 doctors 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 groups 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 2.3; the closer this similarity is to 1, the larger the overlap between social groups in consecutive periods. Figure 17(b) plots a CDF of the average Jaccard similarity for the 5000 ego-centric 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%.

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 does each neighbor appear. We then sort the neighbors in decreasing order of appearance frequency. We compare the first, second and tenth most frequent neighbors to determine if there are groups of neighbors that appear more often and what would be a typical size of such groups.

Figure 18 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. 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 Discussion and conclusion

Our analysis of the Orange mobile network dataset indicates that the usage patterns in Côte D'Ivoire differ drastically from the typical cellphone usage in countries from the western world [10]. Due to the lack of weekly or monthly utilization pattern, we hypothesize that the network utilization is not shaped by people's daily routines, rather peaks in utilization coincide with events in the country.

96% of the territory of Côte D'Ivoire can be categorized as Rural based on the population density [3]. 51.2% of the country population lives in Rural areas while 48.8% lives in Urban areas. At the same time the number of antennas deployed in Rural areas is slightly lower than these deployed in Urban. About a fourth of all the network activity is

initiated by Rural areas, a half of the network activity comes from Urban residents and the remaining fourth is a mix of Suburban and traffic that cannot be identified (due to lack of antenna ID information). This lower activity of Rural compared to Urban residents can be attributed to one of two factors: (i) the population coverage by mobile phone networks is lower in Rural than in Urban areas and (ii) people in Rural areas have lower purchasing power to afford to use mobile communication services. We observe higher number of conducted calls in all scenarios where the mobile call originators and terminators are within close proximity, which indicates high locality of interest in mobile phone communications in Côte D'Ivoire. These results make a strong case for the need of an alternative solution for local voice communication such as [4].

In our evaluation of the relationship between population density and mobile phone usage in Section 3.3, we were surprised to find erratic usage corresponding to very sparsely populated areas, with mean values oscillating between ten and two mean calls per hour. This was in contrast to the more consistent mean number of outbound calls associated with higher population densities. We attribute this to two related factors: sparse antenna placement in sparsely populated areas and antenna placement coinciding with transportation infrastructure. Figure 7 illustrates the coincidence of antennas and the major road system. Even though they are located in regions with low population density, antennas placed alongside major roads transmit a higher number of outbound calls than antennas placed in sparsely populated areas away from infrastructure. Although 42% of recorded antennas are associated with areas classified as Rural, over 96% of square kilometers comprising Côte D'Ivoire's surface area is classified as Rural. With a disproportionate few antennas representing areas of low population density and those antennas behaving in observably different manners based on proximity to infrastructure, it is unsurprising that call duration and frequency are so erratic when observed across areas of low population densities.

We leverage the ego-centric social graphs dataset to analyze the persistence of social groups with which an individual communicates. Our results indicate that two types of social groups exist in the provided ego-centric graphs: (i) social groups that likely occur once and can be attributed to communication activity such as scheduling an appointment, and (ii) social groups that occur persistently over time, which can be associated with regular communication with other subscribers. While the overlap of such persisting social group over time is not very large, there are one to two individuals for each subscriber that persist over the observed period.

Our first hand experience in rural Zambia indicates that often health care initiatives are jeopardized by the lack of reliable information channel between health care providers and targeted individuals. Thus, advanced mechanisms for information dissemination in the context of health care will help significantly improve these services in rural areas. The trends we discover in social groups persistence can serve as a basis for development of algorithms for selection of information relays in egocentric social networks. We hope that the knowledge obtained from such analysis can be further incorporated in information dissemination mechanisms in cases where the ego has limited or no access to a cellphone. We note, however, that while these results are encouraging, further analysis of social groups is needed. Such analysis should focus on social trends in Rural areas specifically and needs to incorporate more information related to individuals' location as well as direction, frequency and duration of calls.

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] European Commission urban-rural typology. `http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Urban-rural_typology`. Accessed: 09/02/2013.

[3] Rural population in cote d'ivoire. `http://www.tradingeconomics.com/cote-d-ivoire/rural-population-wb-data.html`. Accessed: 03/02/2013.

[4] A. Anand, V. Pejovic, E. M. Belding, and D. L. Johnson. VillageCell: Cost effective cellular connectivity in rural areas. ICTD, Atlanta, Georgia, March, 2012.

[5] 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.

[6] R. Chaudhri, G. Borriello, and W. Thies. FoneAstra: making mobile phones smarter. In *NSDR*, San Francisco, CA, 2010.

[7] M. de Bruijn, F. B. Nyamnjoh, and I. Brinkman. Mobile phones: The New Talking Drums of Everyday Africa. 2009.

[8] J. Donner. The rules of beeping: exchanging messages using missed calls on mobile phones in sub-saharan africa. *International Communications Association*, 2005.

[9] 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.

[10] G. Krings, M. Karsai, S. Bernhardsson, V. Blondel, and J. Saramki. Effects of time window size and placement on the structure of an aggregated communication network. In *EPJ Data Science*, May 2012.

[11] A. Kumar, J. Chen, M. Paik, and L. Subramanian. ELMR: Efficient Lightweight Mobile Records. In *MobiHeld*, Barcelona, Spain, 2009.

[12] 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.

[13] M. Minges. Mobile cellular communications in the southern african region. *Telecommunications Policy*, 23(7):585–593, 1999.

[14] S. K. Neil Patel and T. S. Parikh. An asymmetric communications platform for knowledge sharing using cheap mobile phones. In *ACM Symposium on User Interface Software and Technology (UIST)*, Santa Barbara, CA, October 2011.

[15] M. Paik, J. Chen, and L. Subramanian. Epothecary: cost-effective drug pedigree tracking and authentication using mobile phones. In *MobiHeld*, Barcelona, Spain, 2009.

[16] T. Parikh, N. Patel, and Y. Schwartzman. A survey of information systems reaching small producers in global agricultural value chains. In *International Conference on Information and Communication Technologies and Development*, Bangalore, India, December 2007.

[17] 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.

[18] A. Reda, S. Panjwani, and E. Cutrell. Hyke: a low-cost remote attendance tracking system for developing regions. In *NSDR*, Bethesda, MD, 2011.

[19] C. C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.