Characterizing Mobility Patterns of People in Developing Countries using Their Mobile Phone Data

Kuldeep Yadav, Amit Kumar, Aparna Bharti, and Vinayak Naik

Indraprastha Institute of Information Technology (IIIT), Delhi, India
kuldeep@iiitd.ac.in

***Abstract*—Location data collected from mobile phone users provide an ideal platform to generate human mobility patterns. These patterns give us insights into how people travel in their day-to-day lives. With availability of cellular data, either at large-scale but with low location accuracy or at small-scale but with high location accuracy, studying mobility patterns is now possible. An example of former dataset is CDRs (Call Detail Records) and that of latter is GSM/WiFi/GPS traces collected from mobile phones. So far the studies have been focused on data collected in developed countries.**

In this paper, we make a first attempt in finding and analyzing mobility patterns of people in developing countries using both the categories of data. We use publicly available CDR data and we collect our own data for capturing fine-grained location. Ours is the first dataset of its kind that is publicly available. We analyze this data to find movement as well as place visiting patterns, compare our findings with existing studies, and discuss their implications. For example, urban people in developing countries travel farther distances in their day to day life as compared to people living in non-urban areas. Also, distance travelled by urban people in developing countries is as much as six times lower compared to developed countries.

I. Introduction

Location has been an integral part of an individual's context because it can be used to infer several key attributes of her mobility, i.e., places that she visits [5], frequent traveling routes, interactions with other people [9], etc. The number of mobile phone subscribers have reached 6.8 billion covering nearly 95% population of the world [1]. Today's mobile phones provide an ideal platform to find and understand patterns in human mobility. There are primarily two sources of location data collection using mobile phones, collecting data at the cell tower and using an application running on the mobile device. In case of cellular network, identifier of a cell tower (popularly known as Cell ID) is collected as part of Call Detail Records (CDRs), when a phone connects to the network to make or receive a phone call, send or receive a SMS, or has an active data connection [11], [15], [16]. On mobile phone, there are various location interfaces, GPS, WiFi, and GSM [8]. These location interfaces provide different level of accuracy and availability. For instance, GPS provides fine-grained location of a person but do not work in indoor environments [21].

Both the primary sources of location data have their own trade-offs. CDRs collected from cellular network provide an opportunity to perform analysis to find large-scale mobility patterns. Such large-scale analysis can not be performed when data is collected from individual's mobile phones [10], [9]. However, there exist significant amount of flexibility while collecting location from individual mobile phones such as location sampling interval could be set to be high and there is less chance of missing a place, unlike that in CDRs. Similarly, high sampling of location information helps in unveiling the place-visiting patterns as well as accurately estimating place-stay duration. Location studies performed with fine-grained location data offer insights into person-specific mobility [9], [5] but they fail to give broader mobility patterns, i.e., at a scale of city or country, due to limited number of participants.

Related Work: There has been research on using CDRs to study human mobility. Gonzalez et al [3] found that human mobility is highly redundant in spatial as well as temporal dimension. Their work focussed on modeling an individual's mobility pattern using a CDRs dataset of about $1,00,000$ mobile users. CDRs have also been used in diverse application scenarios such as forecasting socio-economic trends [16], characterizing urban areas [17], characterizing human mobility patterns [15], and studying disease spread [4].

In this paper, our focus will be on research related to characterization of human mobility. The characterization factors are (a) Mobility in terms of distance and (b) Mobility in terms of places visited. Place is one level of abstraction above that of distances and associated with other contextual attributes such as frequency of visits, arrival time, and departure time. There has been previous work in characterizing mobility in these two factors. Isaacman et al [13], [11] analyze daily travel of people living in two US cities using a metric *daily range*, which represents the maximum limit of distance travelled by a person in day. Their work reveal several interesting patterns, e.g., people in the one city travel two times more as compared to those in another city on weekdays. Isaacman et al also found considerable difference in people's movement across weekdays and weekends, as well as across different months of an year.

A mobile user is expected to visit several places in a given duration. Researchers have worked on finding places using CDR data [14] as well as fine-grained data collected using individuals' mobile phones [9], [6]. Issacman et al [14] built algorithms to identify important places in a person's mobility history using CDR data. Using ground truth derived from a few volunteers, they found that a person's locations including "Home" and "Work" could be estimated with an error of about 1 mile. Using GPS data, Do et al [5] found out that most people visit $2-4$ places every day and calendar (day/time) has significant impact on people's pattern of visiting places.

Main Contributions: Our contributions are two-fold. One is in analyzing CDR and fine-grained location data for developing countries for mobility characterization. In fine-grained location data, we look at WiFi and GSM location interfaces. We have collected our own dataset for these interfaces, which is now available online. Second is in comparing the mobility characterization with the publicly available dataset from developed countries. While comparing, we not only use available characterization on existing data but also perform our own characterization for newly released data.

To the best of our knowledge, there is no such study, which finds movement and place visiting patterns of mobile users in developing countries. We hypothesize that human mobility patterns in the developing countries be different from those in the developed world due to differences in quality of transportation, socio-economic status, and population density, etc. In this paper, we performed a detailed characterization of human mobility in a developing country using a CDR dataset and compared our findings with similar studies performed in US cities [13], [11]. Some of the questions, which we answer through our analysis, are as follows.

1) What is typical distance and maximum distance travelled by people in their day-to-day lives and how does user demographics impact the distance travelled?
2) Which day of the week is preferred for long distance travel?
3) How many places are visited typically by mobile users and how many of these places are regularly visited?

Our analysis using CDR dataset revealed that there is a significant impact of people's demographics on their mobility, for instance, urban people travel farther distance than their non-urban counterparts. We found that both urban and non-urban people have limited movement on weekends. In comparison to existing studies in developed countries, we observed that distance travelled by urban people in Ivory Coast is as much as six times lower compared to that for urban people in US.

There is a tradeoff in finding human mobility patterns from different data sources but we believe that they offer complementary insights into human mobility. In addition to answering questions and finding patterns from CDRs dataset, we have used fine-grained data collected from individual's phones to answer the following questions. We compare answers to the following questions across two different countries of India and Switzerland.

1) What is typical number of places visited by people per day as well as in total? How frequently do people visit these places?
2) What is a tradeoff in discovering places using GSM and WiFi?
3) How many new places are visited by people and how much time she is likely to spend at the new place?
4) What is a preferred day and time to visit these new places?

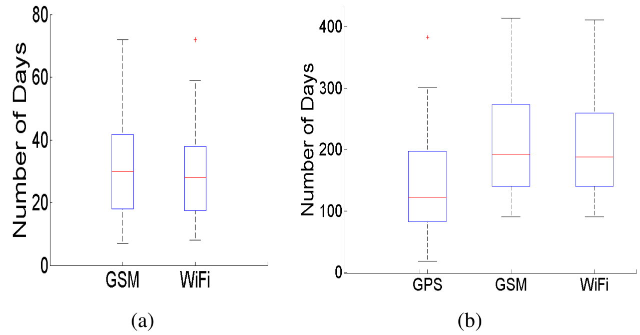

Fig. 1: (a) Data collection days of different location interfaces for all the participants in *dataset 2*; (b): Data collection days of different location interfaces for all the participants in *dataset 3*.

We observed that GSM-based place inference is prone to merge places as compared to that using WiFi-based. However, GSM provides a wider coverage, especially in developing countries, as WiFi is not available at many places. We found that most of the people in developing countries visit lesser number of places as compared to those in developed countries. The most of the new places are visited by participants only for short durations and there is a difference in preferences with respect to time of the day.

II. Datasets

We have used following datasets to find mobility patterns.

1) **Dataset 1 (Ivory Coast)**: This is a CDR dataset, acquired as part of Orange D4D challenge[1]. This dataset was collected in an African country of Ivory Cost by one of the biggest mobile operators Orange. The original dataset contains 2.5 billion records, calls, and text messages exchanged between 5 million anonymous users during a period 5 months from December 2011 to April 2012. We analyze the mobility traces of $50,000$ people, for whom there are Cell IDs in the CDR. Although the complete duration of this data is 20 weeks, every two weeks there is a change in user IDs to preserve privacy of mobile users. Unless and until specified, we have used the dataset of first two weeks for our analysis. However, we have found that our findings were consistent across all the different periods of 2 weeks in this dataset.
2) **Dataset 2 (India)**: This is a self-collected dataset[2] in New Delhi, India. We have developed a data collection

[1] http://www.d4d.orange.com/home
[2] http://muc.iiitd.edu.in/datasets/home.php

tool for Android phones and deployed among 62 participants during March 2012 to November 2012. The participants included graduate and undergraduate students and university technical and administrative staff. The participants were selected using convenience sampling and only criteria used for recruitment was availability of Android phone. Data connection costs were covered of all participants for the whole duration of data collection. Our data collection tool scans and logs timestamp, MCC, MNC, LAC, Cell ID, and RSSI information every 1 minute. Every 10 minutes, it scans visible WiFi APs and logs their SSID and BSSID information with timestamps. We have kept high scanning intervals for WiFi to reduce energy consumption [21]. Mobile data collection tool provides user with an option to automatically sync collected location information to the cloud at any interval between five to thirty minutes.

Our data collection tool collected about 11 million GSM records and about 1 million WiFi records. Spatial diversity of the collected data was high, participants encountered 11847 unique Cell IDs and 7717 unique WiFi APs in the whole duration of the data collection. We have computed the number of days for which data of different location interfaces was collected for every participant as shown in Figure 1a. In case of GSM, nearly half of the participants uploaded data for about 30 days, whereas in case of WiFi it was 28 days.

3) **Dataset 3 (Switzerland)**: It is a public dataset, which was released as part of Nokia Mobile Data Challenge (MDC) 2012 [19], [20]. This dataset was collected in Switzerland from 2009 to 2011 using Nokia $N95$ smartphones. Although original dataset was collected with 200 participants, they have publicly released data of only 38 participants. Dataset contains continuously collected mobility (GPS, WiFi, and GSM), social interactions (Call, SMS, and Bluetooth), and application usage data for all the participants. We have considered only mobility data for our analysis in this paper.

In total, this dataset had about 80 million GSM records, 28 million WiFi records, and 15 million GPS records. GSM information was scanned every 1 minutes, WiFi scanning was performed every 2 minutes, and GPS coordinates were sampled every 10 seconds. Similar to *dataset 1* , spatial diversity of the dataset was also high due to large duration of data collection and large number of participants. There are 18321 unique Cell IDs and $1,26,968$ unique WiFi APs. As shown in Figure 1b, dataset had about 122 days of GPS data, 191 days of GSM data, and 188 days of WiFi data for nearly half of the participants.

III. Mobility Characterization in Terms of Distances

In this section, we analyze aggregated movement patterns in terms of distances across three datasets to answer questions including typical and long distance travelled by participants and preferred days of long distance travel.

A. Daily Ranges

One of the metrics to measure human movement is *daily range*, which represents the maximum distance traveled by a person in day. For instance, if a person visits locations $\{C_1,C_2,C_3,....C_k\}$ in a day then the *daily range* will be the maximum pairwise distance between these locations, i.e.

$$dailyrange(d) = maximum(distance(C_i, C_j)), \quad where \forall i, j \in (1, k) \tag{1}$$

Isaacman et al [11] used *daily range* to find lower bound of a persons' travel and verified measurements of *daily range* with the ground truth provided by volunteers. Median daily range for a person represents the most frequent travel, e.g., home to workplace or vice-versa. Maximum daily range represents the occasional long distance trips that she undertakes, e.g., short-vacations on weekends.

1) Dataset 1: Ivory Coast has 255 subprefectures. We found top 3 subprefectures, which have the highest population density per square mile among all the subprefectures. Our assumption is that an urban subprefectures will have higher population density in comparison to others. Based on this, we divide the total number of participants into two distinct parts, participants living in urban areas termed as urban participants and rest of the participants termed as non-urban. We computed percentiles of median and maximum daily ranges of $37,158$ non-urban participants and $12,842$ urban participants for weekdays and weekends separately as shown in Table I. Here are our findings.

- In case of non-urban participants, on weekdays, $50th$ percentile of median daily range is zero that represents that more than half of participants did not travel in a day to day scenario. Similarly, $25th$ percentile of maximum daily range is zero on weekdays, which represents that nearly one fourth of non-urban participants did not move in the duration of 2 weeks. On weekends, $50th$ percentile of maximum daily range is zero, which represents that majority of non-urban people in developing countries have limited movement on weekends as compared to weekdays.
- Higher value of $95th$ percentile of maximum daily ranges suggests that some (˜5%) non-urban participants undertake occasional long distance travel on both weekdays and weekends.
- We have found significant differences in people's movement across non-urban and urban participants. In case of urban participants, all percentile values of daily ranges are at least 1.5 times higher as compared to their non-urban counterparts. Also, $95th$ percentile of median daily range in case of urban participants are at least 30 times higher on weekdays and 18 times higher on weekends. This findings show that some of urban participants (˜5%) undertake long distance travel on a daily-basis.

- Further, we have observed that movement of urban participants remains restricted on weekends, similar to their non-urban counterparts and few of them undertake long distance travel on weekends.

To summarize our findings, we have found that demographics have significant impact on people's travel i.e. urban people in Ivory Coast travel farther distances in their day to day life as compared to people living in non-urban areas. Also, we have found that most of the urban and non-urban people in Ivory Coast have limited movement on weekends, however some of them undertake long distance travel.

2) Dataset 2 *& 3 :* Cell ID information is sampled at every one minute in *dataset 2* and *dataset 3*. However, *dataset 1* is sparse as Cell ID is recorded only when a person makes call or SMS and it is likely to miss some of the location names which are visited by users but did not get recorded. We converted Cell IDs into corresponding geo-coordinates using Google's crowdsourced database [8]. After that, we found *daily range* using Cell ID's geo-coordinates for each participant. Table II presents percentile values showing distribution of daily ranges of all participants in *dataset 2* and *dataset 3* for both weekdays and weekends. Some of our observations are as follows:

1) In case of *dataset 3*, $50th$, $75th$, and $90th$ percentile daily range on weekends are higher as compared to weekdays, which suggests that people tend to travel long distances during weekends and this is higher than the *dataset 2* too. This observation suggest that on weekends, majority of people in Switzerland (*dataset 3*) are likely to travel, while it is very less in India (*dataset 2*).
2) There is a significant different in daily ranges across two datasets, for instance, $50th$ percentile in *dataset 3*is 2.6 times higher on weekdays and 4.2 times higher on weekends as compared to *dataset 2*.

To summarize, people in India move lesser distance in their day to day life as compared to people in Switzerland. Also, majority of people in Switzerland go for larger distance on weekends as compared to weekdays, where as majority of people in India travel less on weekends compared to weekdays.

Datasets	Dataset 2		Dataset 3	
Percentile	Weekday	Weekend	Weekday	Weekend
5	0.26	0.18	0.95	0.41
25	0.95	0.48	2.49	2.03
50	1.94	1.30	5.09	5.44
75	9.88	6.78	17.61	22.33
95	21.96	25.93	39.74	60.85

TABLE II: Percentile values of daily ranges across all participants and days in *dataset 2* and *dataset 3*.

B. Long Distance Travel

We observed that few urban and non-urban people in *dataset 1* travel large distances on both weekdays and weekends. We are interested in finding out which day of the week, people are more likely to do long distance traveling. We define long distance travel as the maximum non-zero daily range achieved by a participant in the given time period of two weeks. For every day of the week, we count the number of urban and non-urban participants, who achieved maximum daily range. We ignored the participants, for whom daily range was zero in the complete duration.

1) Dataset 1: This dataset is divided into 10 different time periods, we compute average for every day of the week across these time periods. As shown in Figure 2, we found that the most of urban and non-urban participants do their travel either on a Friday or Saturday, while latter being the most preferable day. Sunday was the least preferred day for long distance travel in Ivory Coast.

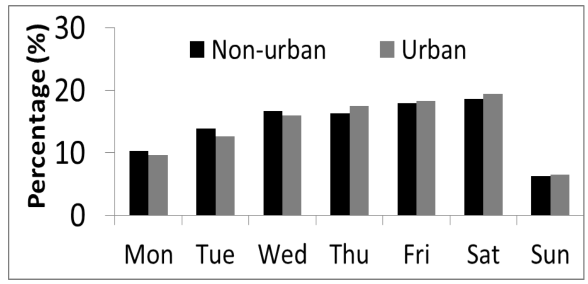

Fig. 2: Weekdays vs average number of people who preferred traveling long distance. Saturday was the most preferred day for long distance travel.

2) Dataset 2 & 3: Due to limited number of participants in *dataset 2* and *dataset 3*, we do not analyze participant specific travel. Instead, we find which weekdays are preferred by the participants for long distance travel. Long distance travel could be participant-specific, we define long distance travel as the travel, which is more than $75th$ percentile of a participants' daily ranges for this analysis. Figure 3 shows percentage of daily ranges which are termed as long distance travels among different days of the week. *Most of the long distance travel happens on Monday and Friday in India (*dataset 2 *) where as Saturday and Sunday were the two most preferred days of long distance travel in* dataset 3 *(Switzerland).*

One of the implication that participants in India perform their long distance travel on weekdays i.e. Monday and Friday could be that majority of participants were students and they travel to their nearby home towns on Friday evening and return on Monday morning. In case of Switzerland, most people like to travel long distances on Saturday or Sunday as they are likely to have holidays on those days.

C. Comparison with Existing Studies

We compare the daily travel ranges of urban people in Ivory Coast with similar studies done in two urban areas of US [11]. From our analysis in *dataset 1*, we have found that daily travel range of urban people in Ivory Coast is significantly lower than daily travel ranges reported in US cities. A significant number of Ivory Coast people do not travel in their regular days ($25th$ percentile is zero for median daily travel range), while in all the US cities, $25th$ percentile was at least more

Datasets Percentile	Dataset 1 (Non-urban Participants) WD-Median	WD-Max	WE-Median	WE-Max	Dataset 1 (Urban Participants) WD-Median	WD-Max	WE-Median	WE-Max
5	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0.96	0	0
50	0	3.89	0	0	0.63	5.24	0.50	1.14
75	1.27	19.53	1.38	6.49	2.39	465.04	2.28	9.13
95	15.66	474.29	24.34	454.74	463.14	469.30	462.43	467.92

TABLE I: Median and maximum daily ranges (miles) computed from trajectories of $37,158$ non-urban and $12,842$ urban participants in *dataset 1*. WD-Max represents the maximum daily range on weekdays and WE-Max represents maximum daily range on weekends.

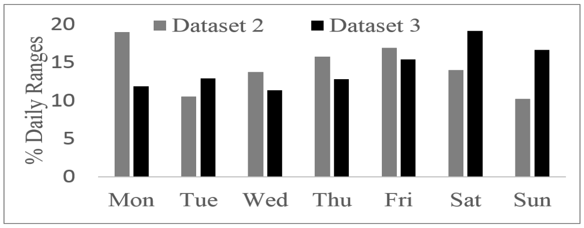

Fig. 3: Weekdays vs average number of people who preferred traveling long distance. Saturday was the most preferred day for long distance travel.

than a mile. Similarly, $50th$ percentile of median daily range for US is at least 6 times as that of urban parts of Ivory Coast. The postulates for limited daily travel ranges in Ivory Coast are as follows.

1) Cell tower density is sparse in Ivory Coast as compared to that in US. Participants may be visiting places but they do not get recorded due to absence of CDR events.
2) The places, to which participants travel, may be close to their home and Cell ID do not change. For instance, many people may be living near to their workplace.
3) There may be significantly high number of participants, who are housewives and they do not travel regularly.

One of the similarities between Ivory Coast and US is that movements of participants are limited on weekends. However, we found differences w.r.t. long distance travel and day of the week. In two US cities [11], researchers found that people prefer to travel long distance on weekends (i.e. Saturday and Sunday) and they have also considered Friday as a part of weekend because a large number of people do long distance travel on that day as well. Similarly, weekend days were preferable for long distance travel in Switzerland too. In our analysis, we observed that Friday and Saturday are indeed the two most preferred days for doing long distance travel in Ivory Coast with a notable exception of Sunday. Sunday was the least preferred day for long distance travel in Ivory Coast as well as India.

IV. Mobility Characterization in Terms of Places

A person visits different places in a day and it is feasible to automatically discover most of these places from their mobility data. We have different modalities of location data in the three datasets. They are CDR, GSM, WiFi, and GPS. The Cell IDs present in CDR and GSM can be geo-located using publicly war-driven databases. One example is www.opencellid.org Apart from CDR, we consider GSM and WiFi data for place discovery as they exist in both *dataset 2* and *dataset 3*. A geo-location is not a place. We need an algorithm that can cluster geo-locations into places.

A. Clustering Algorithms for Each Modality

CDR-based and WiFi-based clustering are well studied problems. Relatively, GSM-based clustering is less studied. This paper is about finding patterns in places visited by people from three datasets and comparing those patterns and not proposing new algorithms for clustering. Hence, we briefly discuss best-known algorithms for CDR-based and WiFi-based and our own algorithm for GSM-based.

1) **CDR-based Clustering** : Previous studies have shown that a user's phone may connect to different cell towers even if she stays at the same place [18]. We implement the algorithm presented by Isaacman et al [14] for clustering Cell IDs. The algorithm uses Hartigans leader algorithm to cluster nearby Cell IDs with the help of a threshold distance (t_d). This algorithm takes into account all the Cell IDs observed by a person in a given time period.
   An assumption in this algorithm is that all the places are at least t_d distance away from each other. Isaacman et al found that value of t_d equal to 1 mile works well in the dataset, which was collected in urban areas of two US cities. To find a good value of t_d, we performed an experiment, where we varied the value of t_d from 0.5 mile to 5.5 mile and computed average number of clusters for $50,000$ users. As it is seen from Figure 4, average number of clusters nearly remains same, if value of t_d is equal or bigger than 1.5 mile. We selected the value of t_d equal to 1.5 miles for our experiments.
2) **WiFi-based clustering** : An assumption in WiFi-based clustering is that a person will observe different set of WiFi APs in different places. We have used UIM clustering algorithm presented in [9], which clusters WiFi APs into a set of distinct places.
3) **GSM-based clustering**: Unlike CDR data, Cell IDs are logged at a fine-grained interval (nearly 1 minute) in a mobile phone and it is less likely to miss a place. However, clustering of Cell IDs has its own challenge as Cell ID may change even when a user stays at same

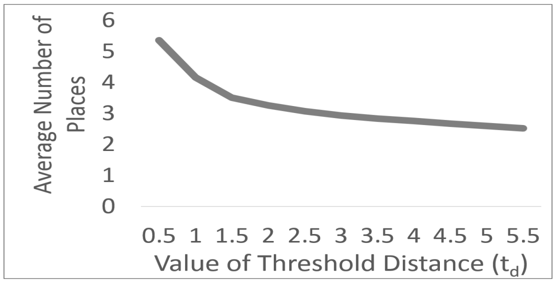

Fig. 4: Effect of changing t_d on the average number of places.

place due to network load, small time signal fading, and inter-network ($2G$ to $3G$ or vice versa) handoff. Such a change in Cell ID while the user is stationary is called as "oscillating effect" [18], [6]. To solve the challenge, we use a graph-based clustering algorithm (GCA) described in one of our earlier work [6]. GCA models the oscillating effect among Cell IDs using an undirected weighted graph (movement graph) and then performs clustering with the help of heuristics such as edge weights, node degree, etc. The evaluation on two diverse datasets show that GCA was able to discover about 80% of the actually visited places from ground truth.

B. Aggregated Place Visits

It is likely that number of visited places are different for every person. However, we are interested in analyzing aggregated place visiting patterns across all three datasets by applying algorithms described earlier.

1) Dataset 1 *:* We applied CDR clustering to find total number of places visited by participants in *dataset 1*. Due to sparseness in CDR data, it is not feasible to discover places visited by a person on a daily basis. Therefore, we applied CDR clustering algorithm to complete data of each participant to find places. As shown in Figure 5, about 29% of participants visit only one place in the whole time period. Large number of participants (about 67%) visited at the most 3 different places. Some participants visit unusually high number of places. For instance 6% of participants visited more than 10 different places in duration of 2 weeks.

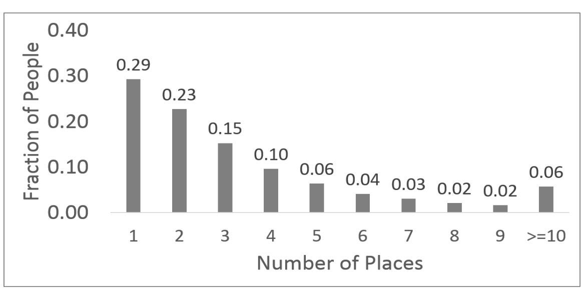

Fig. 5: A histogram showing total number of places visited by participants in *dataset 1*.

Regularly Visited Places: Although a subset of people may visit many places, they may not visit all of these places regularly. The most regular places for a person are likely to be "Home" and "Workplace", where she spends significant amount of time as compared to the places, which are occasionally visited. Hereby, we define a metric *place support value*, which measures regularity of a place in a user's mobility. For instance, *support value* of a place P_i for a given user U_k is computed as:

Support value (U_k, P_i) *= Number of days on which* P_i *was visited by* U_k*/Total number of days for which data is available for* U_k

After discovery of places, we compute a support value for each place. A place is said to be *regular* for a person if its support value is higher than a threshold (δ). If value of δ is equal to 0.3, it means that a place is visited equal to or more than one third of the total number of days. Figure 6 shows the histogram of number of participants with respect to number of regular places, when value of δ is fixed to 0.3.

Majority of users (approx 91% *) had at the most two regular places in their mobility profiles. For some users, we did not observe any regular place. The reason could be due to limited location events in CDRs. Comparing Figure 5 with Figure 6, we conclude that while some users may visit large number of places in a give duration, their regular places such as home and office remain a few.*

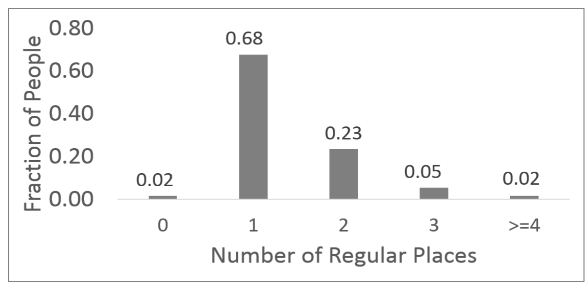

Fig. 6: A histogram showing number of regular places visited by users. Majority of users (91%) had at the most two regular locations in *dataset 1*.

2) Dataset 2 *&* 3 *:* We find places visited by all the participants per day using these datasets. Figure 7a shows the distribution of places visited by participants across all days for two different location interfaces of GSM and WiFi. On majority of days (83%), participants visit at the most two places, where as in nearly 14% days, they visit $3-4$ places every day in *dataset 2*.

GSM vs WiFi-based Clustering: *In case of* dataset 3*, we observe a large difference in distribution of places visited per day among WiFi and GSM. This difference comes from the fact that GSM-based place discovery algorithm merges places, which are near to each other due to longer range of cellular towers.* WiFi-based place discovery approach is able to find two or more different places on different floors of the same building. This level of granularity is not possible using GSM-based approach. For example, library and academic building in a university campus will be situated near to each other. Though WiFi-based place discovery algorithm will discover them as two different places primarily due to limited range of WiFi APs, GSM-based clustering approach may merge them and show it as a single place.

WiFi penetration is limited in developing countries and many places visited by participants in *dataset 2* do not have significant WiFi deployment. According to our estimates, WiFi was available nearly 60% of time in the given data collection duration for *dataset 2* , where as WiFi availability was over 90% in *dataset 3* . Hence, we did not see a difference in places discovered by WiFi and GSM-based approach in the case of *dataset 2*. However, this difference is evident in average number of places visited in whole duration of data collection.

Average # of Places: In case of *dataset 2*, the average number of places visited by the participants were about 12.72 (ρ : 3.83) using WiFi data, where as it is 14.28 (ρ : 7.16) using GSM data. The average number of visited places was higher due to long duration of data collection in *dataset 3* . In particular, 101.31 ($\rho : 50.68$) places were discovered using WiFi data and 65.55 ($\rho : 19.02$) places were discovered using GSM data.

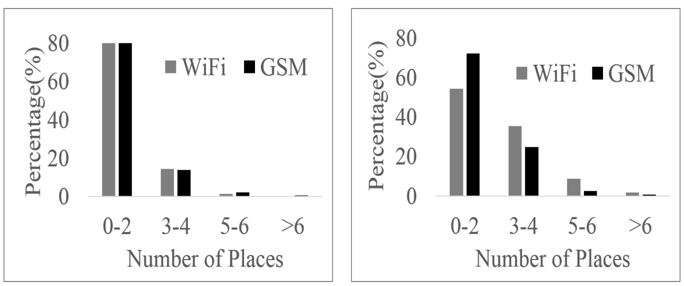

Fig. 7: (a) Distribution of number of places visited by all the participants every day in *dataset 2* ; (b) Distribution of number of places visited by all the participants every day in *dataset 3*.

Typically in a person's mobility history, there are two regularly visited places i.e. "Home" and "Workplace", and that is why, for majority of days, participants were restricted to at the most $1-2$ places in both *dataset 2* and *dataset 3*. However, participants in *dataset 3* visited $3-4$ places on more number of days as compared to that in *dataset 2*. The number of days, where $5-6$ places were visited, is nearly the same across both the datasets as far as data from GSM interface are concerned. *From our analysis, we conclude that people living in Switzerland visit higher number of places per day compared to their Indian counterparts.*

Place Visit Frequency: We find the day count (frequency) of visited places for all the participants in complete duration of data collection. We found that there were some regular places (i.e. "Home") in a person's mobility which were visited often as evident from the long tail of CDF curve in Figure 8a.

Most of the places ($60-70\%$*) were visited only once and nearly* 90% *places were visited less than* 10 *times in whole duration. These findings are consistent across both* dataset 2 *and* dataset 3 .

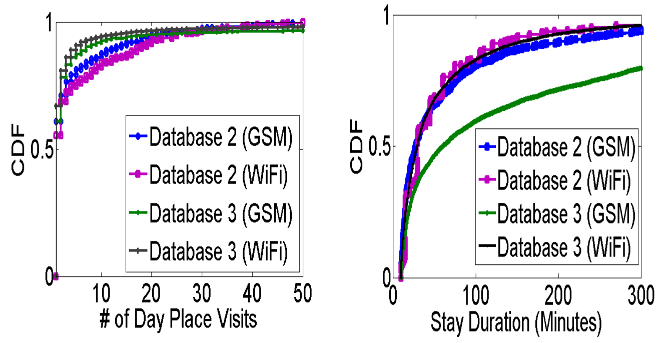

Fig. 8: (a) CDF of total place visits for all the participants in *dataset 2* and *dataset 3* (b) CDF of stay duration at infrequent places for all the participants in *dataset 2* and *dataset 3*.

C. Infrequent Places

People are likely to visit many infrequent (new) places in their day-to-day life. These places may be shopping center, restaurants, friend's home, and holiday spots. Patterns emerging from visits to these places can help in recommendation of advertising. In this section, we find patterns related to infrequent places in *dataset 2* and *dataset 3*. We do not consider *dataset 1* for this analysis as it is likely that many of infrequent places visited by a person may not get recorded by CDR events. For finding infrequent places in *dataset 2* and *dataset 3*, we consider all the places, which have value of place support value less than 0.1. In other words, they are visited on less than 10% of days in complete duration.

Stay Duration: Figure 8b presents the duration of stay for all the participants in infrequent places. While, it is likely that participants spend significantly higher amount of time in regular places, they spend limited time in infrequent places. The median stay duration in case of *dataset 3* using WiFi data was about 29 minutes, where as it was about 60 minutes using GSM data. We believe that this difference in median stay duration between GSM and WiFi data is a reflection of merged places as described in the last section. For *dataset 2*, the median stay duration was approximately 30 minutes for WiFi and 27 minutes for GSM. In case of *dataset 3*, we observed a significant difference in $90th$ percentile of stay duration, 158 minutes in the case of WiFi and 540 minutes in the case of GSM.

Our observations from both the datasets show that the most of the infrequent places visited by participants are only for short duration, i.e., median of the place stay duration is less than 30 *minutes. However, when a person visits farther places, it is likely that she will stay for a longer duration as shown by* $90th$ *percentile of GSM data.*

Place Visit vs day of the Week: As we have observed earlier, regular places are likely to be visited on nearly all the days, where as most of infrequent places are visited only a few times. Figure 9 presents the aggregated view of participants' preferences about different days of a week. *We observed that Saturday was the most preferred day for visiting infrequent places in both the datasets. Participants are less likely to visit infrequent places on working days of week as compared to weekends with notable exception of Friday in* dataset 3 *and Thursday in the case of* dataset 2*. The participants in* dataset 2 *are more likely to visit infrequent places on Sunday than* dataset 3.

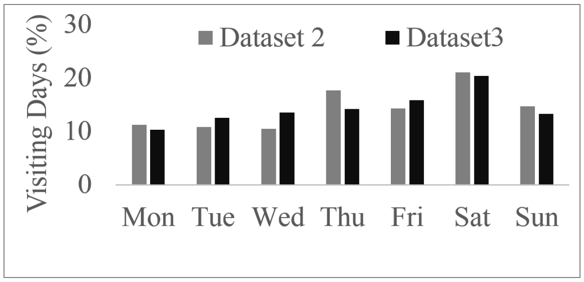

Fig. 9: Histogram showing percentages of infrequent place visits of the participants w.r.t. seven days of week. We have used GSM data of *dataset 2* and WiFi data of *dataset 3* .

Preferable time of the day: We observed earlier that participants spend limited amount of time ($\approx$ 30 minutes) at infrequent places. One of the interesting questions to answer from the dataset is to find if there is any dependency between place visited and time of the day. In other words, at what time of the day participants are most likely to visit infrequent places? Figure 10 shows the percentage of places visited by participants across different time intervals. We use a time interval of one hour each, resulting in 24 different time intervals in a day. *In* dataset 2*, participants are most likely to visit new places during evening time (i.e.* $18:00-19:00$*). However in case of* dataset 3*, participants prefer to visit infrequent places during day time (i.e.* $12:00-13:00$*), which indicates that participants of* dataset 3 *visit new places primarily for lunch.*

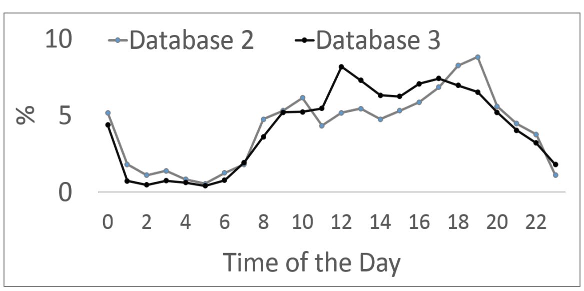

Fig. 10: Percentage of infrequent place visits with respect to time of the day for both the datasets. We have used GSM data of *dataset 2* and WiFi data of *dataset 3* .

D. Comparison with Earlier Studies

Most participants ($\approx$ 65%) visited 4 to 6 number of important (regular) places in the study done in two US cities. In the case of Ivory Coast dataset, most of the participants (68%) remained at one place. These findings show that mobile users in developing countries visit lesser number of places as compared to their counterparts in US. One of the bias in the case of Ivory Coast is that data duration is only 2 weeks as compared to more than 11 weeks data for US dataset. However, when we analyzed different time periods of Ivory Coast data, we found nearly similar distribution for every time period. With a longer duration, number of distinct places may rise but for a typical user's mobility profile, number of regular places are unlikely to change much. After comparing distribution of number of places, our conclusion is that people in developed countries visit more regular places than those living in developing countries, such as Ivory Coast.

V. Conclusion

According to MIT Technology Review [2], building techniques to analyze human mobility patterns from location data collected using mobile phones will be amongst breakthrough technologies in 2013. In this paper, we found human mobility patterns in two dimensions, first using *daily range* and second using places visited by them. Two of our datasets were collected in developing countries of Ivory Coast and India. We have compared mobility patterns emerged from these datasets with studies performed in US and Switzerland. Our analysis and comparisons are guided by set of questions raised in Section I. Our analysis with large-scale CDR dataset showed that people living in urban areas travel lesser distance as compared to people living in urban areas. Similarly, urban people in developing countries travel lesser distance as compared to urban people in developed countries. We observed that both urban and non-urban people in developing countries have restricted human mobility range on weekends, which is consistent with the studies done in developed countries too. A fewer number of people in developing countries were found to be traveling long distance on weekends as compared to those in developed countries. Aggregated movement patterns on a city-scale dataset can assist in making critical policy decisions for variety of domains such as transportation, urban planning, infrastructure development, etc.

We implemented clustering algorithms to discover places from location interfaces of CDRs, WiFi, and GSM data. From our analysis of real-data, we highlighted tradeoff of using different location interfaces for discovering places. For instance, GSM-based approaches merges many places, which are nearby but were able to discover places, where WiFi is not available. In CDR dataset, the number of regular places visited are smaller in developing countries as compared to those in developed countries. In case of place-visiting patterns among participants in India and Switzerland, we found some similarities such as similar median stay duration and Saturday being the most preferred day of visiting infrequent places. Except Saturday, people in India were more likely to visit infrequent places on Sunday, where as in Switzerland it was Friday. The most of new place visits were performed during afternoons in Switzerland, where as those were performed during evenings in India.

REFERENCES

[1] Number of mobile phone subscriptions, http://www.internetworldstats.com/mobile.htm
[2] Big data from cheap phones, http://www.technologyreview.com/featuredstory/513721/big-data-from-cheap-phones/
[3] Barabi A., Understanding individual human mobility patterns," Nature 453, 779-782.
[4] Wesolowski, A. et al. "Quantifying the impact of human mobility on malaria." Science 338, no. 6104 (2012): 267-270.
[5] Do, T et al. "The Places of Our Lives: Visiting Patterns and Automatic Labeling from Longitudinal Smartphone Data." (2013): 1-1.
[6] Yadav, K et al. "MobiShare: cloud-enabled opportunistic content sharing among mobile peers." (2012).
[7] Yadav K. et al,Challenges and novelties while using mobile phones as ICT devices for Indian masses: short paper, NSDR'10.
[8] Yadav, K. et al. Low Energy and Sufficiently Accurate Localization for Non-Smartphones. In Mobile Data Management (MDM), 2012 IEEE 13th International Conference on (pp. 212-221). IEEE.
[9] Vu L. et al. Jyotish: Constructive approach for context predictions of people movement from joint Wifi/Bluetooth trace, PerCom 2011.
[10] Chon Y. et al, LifeMap: A Smartphone-based Context Provider for Location-based Services, IEEE Pervasive Computing, 2011.
[11] Isaacman, Sibren et al. "A tale of two cities." HotMobile'10.
[12] Noulas, Anastasios et al. "A tale of many cities: universal patterns in human urban mobility." PloS one 7, no. 5 (2012).
[13] Isaacman, Sibren et al. "Ranges of human mobility in los angeles and new york." IEEE PERCOM Workshops, 2011.
[14] Isaacman, Sibren et al. "Identifying important places in peoples lives from cellular network data." Pervasive Computing (2011): 133-151.
[15] Becker, Richard et al. "Human mobility characterization from cellular network data." Communications of the ACM 2013.
[16] Frias-Martinez, Vanessa, Cristina Soguero-Ruiz, Malvina Josephidou, and Enrique Frias-Martinez. "Forecasting Socioeconomic Trends With Cell Phone Records." (2013).
[17] Vieira, Marcos R. et al. "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics." SocialCom 2010.
[18] Bayir, M.A. et al, "Discovering spatiotemporal mobility profiles of cellphone users," WOWMOM'2009.
[19] Laurila, Juha K. et al. "The mobile data challenge: Big data for mobile computing research." Pervasive Workshop on the Nokia Mobile Data Challenge, 2012.
[20] Carroll A. and Heiser G., An analysis of power consumption in a smartphone. USENIXATC'10.
[21] Gaonkar S. et al, Micro-Blog: sharing and querying content through mobile phones and social participation,MobiSys '08.