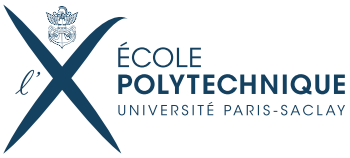
Understanding Mobile Data Demand regarding Mobility

Guangshuo Chen



Mid-Term Thesis Evaluation

Understanding Mobile Data Demand regarding Mobility

Guangshuo CHEN

Ecole Doctorale N.580 STIC
Université Paris-Saclay

June 10^{th}, 2016

Jury Members

Examiner	Marcelo Dias de Amorim CNRS-UPMC Sorbonne Université
Examiner	Lila Boukhatem Université Paris-Sud 11
Supervisor	Aline Carneiro Viana Inria Saclay - Ile de France

Contents

1 Introduction

Smartphones are supposedly the fastest-spreading technology in human history. People live with their phones for both work and personal life. Cellular network, which bears various services and applications, is overloaded by the explosion of data traffic. Global mobile data traffic has a growth of 74% in 2015, and is predicted to have an eightfold increase in 2020 [1]. Hence *the understanding of subscriber's mobile data demand* is of great significance for solutions managing the increasing data traffic as well as improving quality of communication service. Specifically, knowing subscribers' data usage patterns allows telecommunication operators to arrange their network infrastructures against incoming data traffic in advance by using approaches such as data-offloading [2], caching [3] and prefetching [4], and to set mobile data plans having better network resource allocation [5].

A core problem in understanding mobile data demand is *to what degree is mobile data traffic predictable?* In the case of content prediction, [4] studies the repeatability of objects in mobile data content and explores the possible gain of prefetching in mobile network. In the case of data volume prediction, [6] studies dynamics of data traffic of cell and [7, 8] explore the limits of predictability of data traffic of cell. Data traffic of individual is modeled in terms of time in [9], and is analyzed pointing at spatial-temporal correlation in [10] on a small group of subscribers. To the best of our knowledge, there is no work in the literature exploring the maximum predictability of data volume of individuals, or modeling mobile data traffic integrating time and space of large-scale mobile data analysis [11]. The objective of the thesis is (1) *to explore the limits of predictability of data traffic* and (2) *to exploit the 3-dimensional correlation among data demand, mobility, and time* on large-scale mobile datasets.

This paper is the mid-term report of the thesis. We explore the predictability of data volume for individuals. Specifically, our goal is *to determine the maximum probability of forecasting data volume for each subscriber*. To this end, we mine a large-scale mobile dataset with both voice traffic and data traffic, construct a dataset of time series of data volume and explore the upper bound of predictability hidden in the time series. The analysis is carried on using information entropy which is utilized to explore various predictabilities such as mobility [12, 13], road traffic [14] and also data traffic [7, 8]. Besides, in order to understand *whether subscriber's mobility contributes to forecasting data traffic*. We construct a dataset describing mobility and data traffic and summarize regular behaviors among data usage by mining and clustering data sessions for each subscriber.

Our contributions are summarized as following. (1) To the best of our knowledge, our work is the first entropy analysis of individual's data volume on large-scale mobile dataset. (2) We find a overall $> 90\%$ of predictability hidden in individual's time series of data volume. (3) We investigate subscriber's data usage pattern in terms of mobility and time (on going).

The context of the report supports the EU CHIST-ERA Mobile context-Adaptive CAching for COntent-centric networking (MACACO) project. It proposes to explore possibilities to extract and forecast the behavior of mobile network users in the three-dimensional space of time, location and interests for deriving efficient data offloading protocols. Using such a protocol, mobile data would be identified and cached at the network edge early, which reduces data costs and offers better quality of service.

The rest of this report is organized as following. Section 2 describes the GranData dataset and the two datasets that we construct. In Section 3, we exploit the predictability of data volume. In Section 4, we investigate the data usage in terms of mobility. Section 5 summarizes the report and declares the next steps of the thesis.

2 Dataset Description

We have access to a long-term country-scale mobile dataset due to our collaboration with the GranData company [15]. The dataset ensures enough amount of subscribers for surveying trend of population and diversity for analyzing individual's usage patterns. The dataset describes voice traffic by call detail records (*CDRs*) and data traffic by data sessions (*sessions*), collected from a major cellular network of Mexico in a consecutive period of 3 months. Each subscriber has a unique anonymized identifier in a CDR or a session. The CDR consists of subscriber's identifier, role (caller or callee), counterpart's identifier, cell's coordinate, call duration and time stamp. The session consists of subscriber's identifier, volume of data and time stamp.

To perform mobility analysis and predictability analysis, the pretreatment on sessions is required. Firstly, note that the sessions does not contain any information about cell. It is infeasible to perform mobility study on sessions due to the lack of motility information. To overcome the problem, we leverage mobility information from voice traffic by extracting locations from CDRs and complete sessions with information of cell by inferring location for each cell. Besides, we filter out some subscribers due to incompleteness of mobility information. Secondly, a subscriber can have a prepaid contract or a postpaid contract and may use her network service 1 day in a month, or almost every day. For generality of the study, it is necessary to have subscriber with similar contracts and activities. Hereby *we construct two datasets based on the dataset above*: Dataset D_1 for predictability analysis and Dataset D_2 for mobility analysis.

2.1 Dataset 1: Time Series of Discretized Data Volume

Dataset D_1: the dataset contains $162,118$ postpaid subscribers, chosen from approximately 1.6 millions given the criteria that they join the network every day. We select only postpaid subscribers, since they are far more active in using their network than prepaid subscribers: 20% of postpaid devices are used every day against only 2% of prepaid devices. As Figure 1(a) shows, more younger users are select than older users under the given criteria, because younger users are more active in using the network [9].

We perform the predictability analysis of data volume based on information entropy. To this end, we make a time series of symbols for each subscriber by segmenting the 3-month period into one hour-long intervals and discretizing the sum of volume of sessions into symbols. The time series describes the subscriber's data usage: each symbol represents data volume generated in one hour. A symbol is assigned every hour, identified as following: when there is no sessions in a certain hour, a symbol "X" is marked representing the subscriber is *idle* in this hour; otherwise, we assign it as the *magnitude* of the sum of volume of all sessions, mathematically as $\text{ceil}(\log_{10}(\sum v_i))$ where v_i represents volume of the session i. In the discretization, all sessions with volume < 1 *kilobyte* are regarded as background traffic and eliminated. In practice, we obtain 8 magnitudes from 10^0 to 10^7 representing data volume from 1 *kilobyte* to 10 *gigabyte*. In D_1 each subscriber has a time series of length $L = 92 \times 24 = 2208$ consisting of $8 + 1$ symbols.

2.2 Dataset 2: Data Sessions with Time, Location and Volume

Dataset D_2: we construct the dataset by leveraging mobility information from CDRs and inferring location for each session for selected $17,366$ subscribers. Subscribers in D_2 are chosen from the area of Mexico City, not like ones who are selected all over the country in D_1. For generality, 4 weeks of CDRs and sessions are selected from the period of 2 months (October and November) ignoring all national holidays as shown in Figure 1(b). We ignore December considering that people may behave irregularly when Christmas is coming.

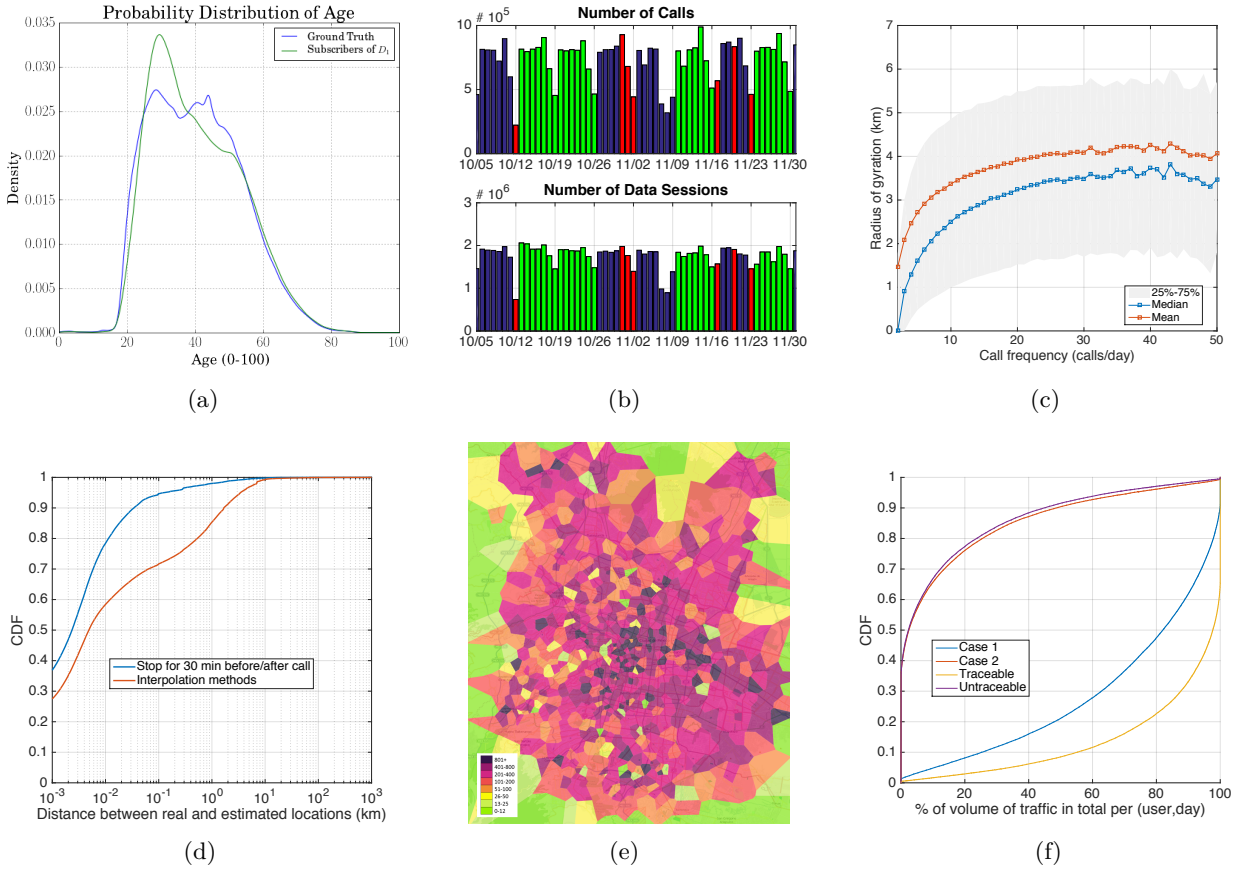


Figure 1: Dataset D_1: (a) Distribution of age of subscribers. Dataset D_2: (b) Number of calls and number of session per day during the two-month-long period (green: selected days, red: holidays); (c) Call frequency $Freq_{call}$ versus median of radius of gyration R_g; (d) Distribution of error of identifying stop locations; (e) Population Density of Mexico City, measured by Voronoi diagram of cells and subscribers' *home* locations; (e) Cumulative Distribution traceable/untraceable volume and sessions in the cases 1 and 2.

Selecting profitable subscribers: not every subscriber who appears in CDRs is appropriate regarding the completeness of mobility information. Locations only appears in CDRs when a call is made or received and some subscribers only make a few calls. Hence we select profitable subscribers by applying the following filters on all subscribers who have appeared in CDRs. (1) Traffic filter: we eliminate approximately 1% of subscribers who generate data < 10 *kilobyte* per day in order to filter out devices generating only background traffic. (2) Mobility filter: the filter is to select subscribers who have enough calls to build mobility trajectories. Explicitly, a subscriber who has ≥ 30 calls is preserved, due to the consideration that if a subscriber has more than 30 calls, it is likely to capture all his passing cells. As shown in Figure 1(c), median of radius of gyration of subscriber R_g increases with the call frequency $Freq_{call}$ per day, but become stable when $Freq_{call} >= 30$, indicating that averagely 30 calls are enough to capture one day's mobility. (3) Outlier filter: we eliminate subscribers who always play the same role (caller or callee) all the day, in order to eliminate devices of call center. Such a device is abnormally utilized, serving an irregular high number of calls from the same direction, which occurs as a typical call center scenario, *i.e.*, a device is used by people for making outgoing calls.

Note that we select subscriber by her behaviors on a daily basis, *i.e.*, on a certain day, if a subscriber is retained after filtering, all her CDRs and sessions of that day are solely preserved. There are 17,366 subscribers with 2,398,392 calls and 954,737 sessions in the dataset D_2. On average, each selected subscriber has 44.7 calls and 17.8 sessions of 29.6 *megabytes* of data per day.

Leveraging mobility information from CDRs: one tends generally to stay in the surroundings of her call places and move rapidly between places of communication [16], which serves as the foundation of mobility extraction based on calls. For each subscriber, we build her trajectory by identifying stop locations and home locations.

(1) *Stop location* is the cell and the time period in which the subscriber's communications are handled by the cell. We assumed that the handover does not happen before/after 30 minutes. In other words, in this time period of $[-30min, 30min]$, all sessions are assumed to be handled by the cell of recent call, used in [12, 10]. If two consecutive calls occur in the same cell within 2 hours, we assume that the cell handles all the communications from the beginning of the earlier call to the ending of the later. Figure 1(d) presents the distribution of error distances between stop location and real location, generated by applying the same approach on 84 subscribers with known GPS positions captured every 5 minutes collected in the MACACO project. We resample positions of these subscribers by following the distribution of CDRs in D_2, and recover missing locations according to time by the approach in [17] for comparison.

In most instances, stop locations are close to subscribers' actual locations.

(2) *Home* is the most active cell of the subscriber during the midnight used in [18, 19]. The activity of cell is measured as the ratio of the total active duration of that cell to the one of all cells, *e.g.*, in $(22h, 7h)$, the ratio of the cell k as $r_{k,22-7} = d_{c,22-7}/D_{22-7}$. For each subscriber, we identify the cell having the largest activity during $(22h, 7h)$ in the 3-month-long period as *home*. Considering that one may have unusual living pattern infrequently, *e.g.*, going to a party instead of home during the midnight, for each day the *home period* is identified as the longest time period during $(20h, 9h)$ in which every call occurs at *home* found. During the home period, all communications are assumed to be handled by the *home* cell. We are able to identify *home* and *home period* for 97.69% of subscribers. Figure 1(e) presents the population density of Mexico City, based on detected home locations.

Infer locations for sessions: for each subscriber, an one-day long period are grouped by two types of time periods: periods of stop locations and home are *traceable* periods; the rest are *untraceable* periods. Every session has the attribute of time t pointing when it happens. We are able to acquire the subscriber's current cell if the communication occurs during a traceable period. For each session, if t is in a traceable period, it could be in the case 1: t belongs to one of stop locations, or in the case 2: t is in the subscriber's home period. Figure 1(f) shows that most of volume are traceable for each subscriber and a traceable session is mostly found in a period of stop location. On average, a subscriber generates 15.1 sessions and 25.1 megabytes of data traffic daily in traceable cases. To conclude, though a few sessions are untraceable, studying subscriber's data behavior in terms of his mobility using only traceable sessions does not include a big bias since most of sessions and data traffic are involved.

3 Leveraging the Predictability among Data Traffic

We exploit information entropy [12, 13] to quantify the uncertainty and the predictability of data volume. Since most of subscribers have data traffic only a few hours per day [9], *forecasting whether a subscriber generates data (is not idle)* is as important as *forecasting data volume*. Hence both idle prediction and volume prediction are considered together. Each subscriber in D_1 has a time series of symbols representing data volume. Since a subscriber has at most 9 unique symbols (**idle** + 8

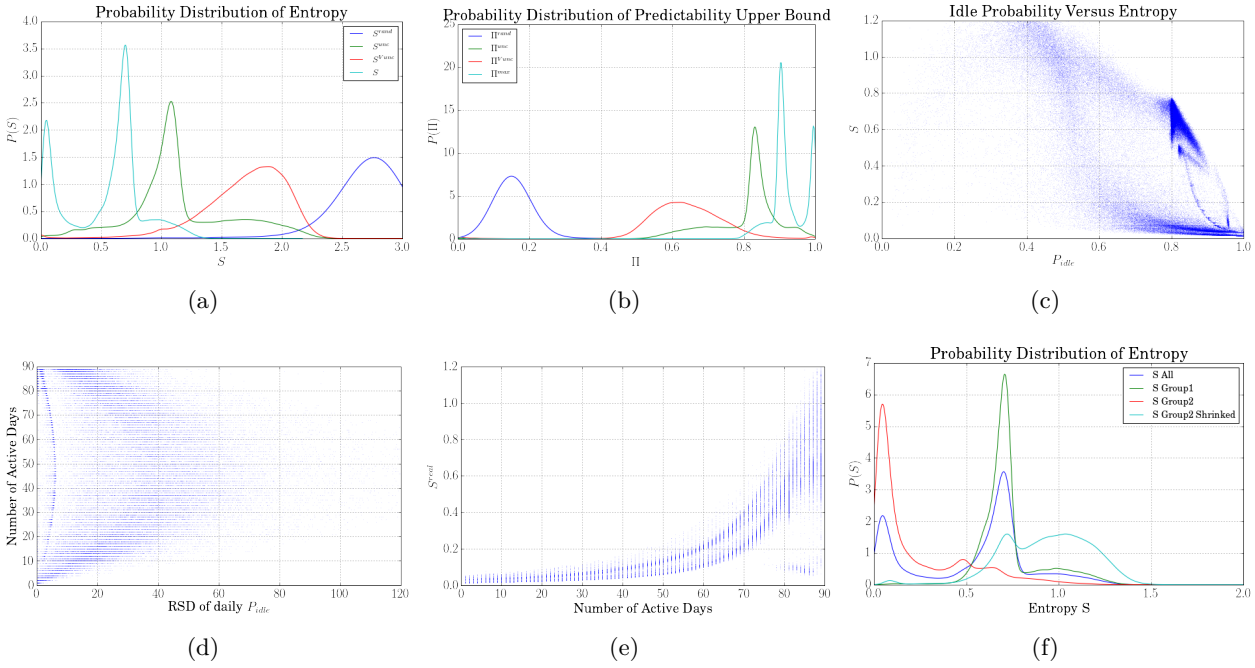


Figure 2: (a) Distribution of entropy; (b) Distribution of upper bound of predictability Π^{max}; (c) P_{idle} versus real entropy S; (d) Relative standard deviation (RSD) of P_{idle} per day versus number of active days; (e) RSD of P_{idle} per day versus real entropy S; (f) Distribution of real entropy of grouped subscribers.

magnitudes of volume), in this case a time series is regarded as result of a stationary stochastic process. We study the predictability of forecasting the proper symbol in each hour-long interval by analyzing information entropy of the time series.

3.1 From Entropy to Predictability

Entropy captures the degree of predictability characterizing a time series [12]. We assign entropy measurements to each subscriber's time series as following: (1) *random entropy*, $S^{rand} = \log_2 N$ where N is number of unique symbols, representing the degree of predictability of a time series where each symbol appears with the same probability; (2) *temporal uncorrelated entropy*, $S^{unc} = -\sum_{i=1}^{N} p_i \log_2 p_i$ where p_i is the probability of symbol i (8 magnitudes + *idle*), characterizing the heterogeneity of data usage patterns; (3) *volume temporal uncorrelated entropy*, S^{Vunc}, same as (2) without *idle* (8 magnitudes); (4) *real entropy*, S, capturing the full spatial-temporal order present in a subscriber's data usage pattern, determined by both the probabilities of symbols and their orders of appearance. We use an estimator for S based on the Lempel-Ziv data compression [13]. Note that $0 < S < S^{Vunc} < S^{unc} < S^{rand}$ [13]. Note that the entropy is in inversely proportional to the predictability, *i.e.*, a time series with lower uncertainty is easier to be predicted.

Predictability is the overall probability of properly predicting the state of volume on each interval for a time series with infinite length, defined as $\Pi := \lim_{n\to\infty} \frac{1}{n} \sum^{n} \Pi(i)$ where $\Pi(i)$ is the predictability of the ith interval. [13] proves that for a time series having N symbols and entropy S, its predictability Π satisfies $\Pi < \Pi^{max}(S, N)$ where the maximum predictability Π^{max} is the solution of the non-linear function $S = -[\Pi^{max} \log_2 \Pi^{max} + (1-\Pi^{max}) \log_2 (1-\Pi^{max})] + (1-\Pi^{max}) \log_2 (N-1)$. For a subscriber with $\Pi^{max} = 0.8$, we can hope to predict her data volume (as magnitudes) and whether being *idle* for

at most 80% of hour-long intervals; for the rest 20%, the symbol appears randomly.

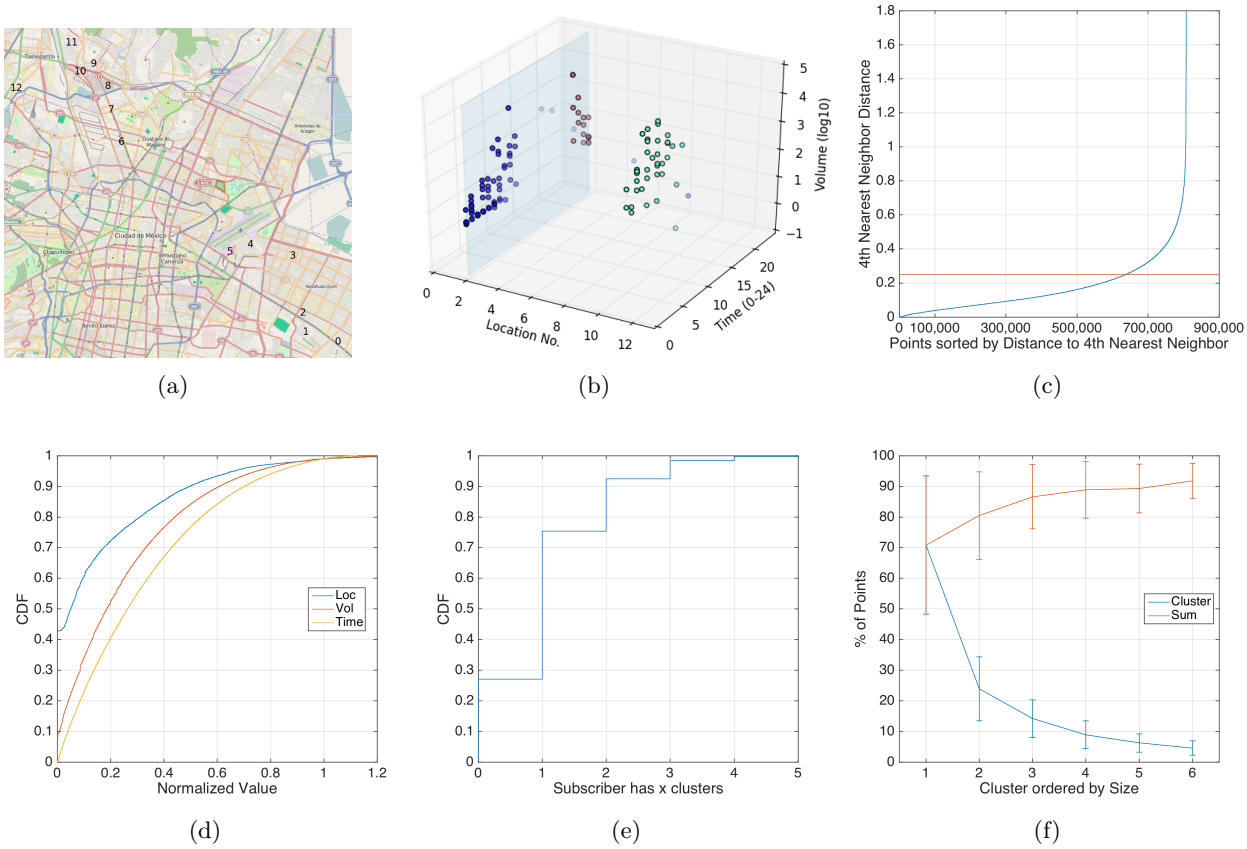
3.2 Maximum Volume Predictability of Subscribers

We determine the maximum predictability for each subscriber in D_1 based on 4 entropy measurements. Π^{max} is determined as the maximum predictability by N and real entropy S. Similarity, Π^{Vunc}, Π^{unc} and Π^{rand} represent the maximum predictability using S^{Vunc}, S^{unc} and S^{rand}, respectively. We observe that $P(\Pi^{max})$ is narrowly peaked near $\Pi^{max} = 0.90$ and $\Pi^{max} = 0.99$, corresponding to the two peaks in $P(S)$, as shown in Figure 2(a) and 2(b). The distribution of $P(\Pi^{max})$ indicates that *the time series of properly symbolized volume contains a high degree of potential predictability* for most of subscribers. $P(\Pi^{unc})$ is peaked at $\Pi^{unc} \approx 0.84$ but $P(\Pi^{Vunc})$ is widely distributed. The difference between two distributions indicates that relying only on the heterogeneous spatial distribution, we can expect significant predictability when forecasting magnitudes as well as *idle* (*cf.*, $P(\Pi^{unc})$), but can only have insignificant predictability varying from person to person when forecasting magnitudes (*cf.*, $P(\Pi^{Vunc})$). In other words, due to the fact that most of subscribers generate data traffic on a few hours in a day [9], the predictability of where being *idle* is always high even without considering temporal order of the series. For forecasting magnitudes of volume, relying on the probabilities of each symbol is not enough. Besides heterogeneous information, temporal information of the series has to be considered.

Real entropy S has two peaks on $S = 0.05$ and $S = 0.75$, indicating the existence of two groups of subscribers. For each subscriber, we identify P_{idle} as the probability of discovering *idle* in an interval. As Figure 2(c) shows, subscribers grouped at $S^{=}0.75$ have $P_{idle} \approx 0.8$ while subscribers grouped at $S = 0.05$ have P_{idle} varying from 0.7 to 1. The uncertainty of the former group is approximately $2^{0.8} \approx 1.74$, indicating that a subscriber may have two most possible symbols (*i.e.*, *idle* + 1). Considering that P_{idle} is the biggest in general, we infer that *a subscriber in this group generates data mostly in only one magnitude of volume* though 8 magnitudes are optional. The uncertainty of the latter is about $2^{0.1} \approx 1$, indicating that a subscriber of this group may always generates data in only one magnitude of volume or do not generate any traffic most of their time. Since most of subscribers in this group have a $P_{idle} > 0.85$, we infer that they rarely make data traffic.

We observe that P_{idle} affects the predictability. A basic explanation is that a time series with lower P_{idle} contains more opportunity of having a huge uncertainty. To determine how much predictability is rooted in the subscriber's active pattern. We compute the relative standard deviation (RSD) of daily P_{idle} by using subscriber's P_{idle} of each day in 92 days. As Figure 2(d) shows, extremely "active" (90 active days) as well as "lazy" ($\leq$ 5 active days) subscribers have RSD $<$ 20 subscribers, indicating both types of subscriber respectively have a similar daily data usage pattern. We observe that for a subscriber, her maximum entropy is determined by number of active days (Figure 2(e)). Hence, we infer that for a subscriber with a few active days, her daily usage pattern is hidden because of having a high P_{idle} every day. To reveal the uncertainty of such a subscriber, we categorize subscribers into two groups: ones making data traffic every day as group 1 and the rest as group 2. For a subscriber in group 2, we also shrink her time series by removing days with $P_{idle} = 1$, named group 2 shrinked. As shown in Figure 2(f), each group has only one narrow peak of $P(S)$, corresponding to the two peaks of real entropy $P(S)$ (Figure 2(a)). More uncertainty are found from shrinked time series, indicating that *for a subscriber who rarely generates data, we can hardly expect a significant predictability of volume.* $P(S)$ of both group 1 and group 2 shrinked is peaked at $S \approx 0.8$. Recalling the discussion above, most of subscribers generate data in only one magnitude of volume.

In summary, *one can expect a high degree of predictability of data volume across the population. It is relatively hard to predict data volume for inactive subscribers.* It is easier to forecast whether being *idle* than magnitudes of volume, which contributes to the high degree of potential predictability. In order to achieve an accurate data volume forecasting, it is important to profile data traffic and characterize



(a) (b) (c)

(d) (e) (f)

Figure 3: (a) Order of locations in (b) and their physical positions; (b) A subscriber's 3-dimensional space of location, time and volume: colors represents sessions' clusters and blue surface represents home location; (c) Distribution of distance to the 4th nearest neighbor: blue line represents sessions sorted by distance and red line represents $\epsilon = 0.25$; (d) Distribution of normalized distance $dist^{(location)}$, $dist^{(time)}$ and $dist^{(volume)}$; (e) Distribution of number of clusters per subscriber; (f) Ratio of sessions that a subscriber's top x biggest cluster account for (x is from 0 to 6).

subscriber by traffic behaviors as in [9]. Besides, it is also of significance to bring context information, *e.g.*, mobility information of subscriber.

4 Investigating Mobility and Data Demand

We analyze *how a subscriber generate data traffic in terms of her mobility* using dataset D_2 in this section. A subscriber in D_2 has certain days captured having enough mobility information, *i.e.*, a majority of sessions have known locations from a trajectory of locations covers most of the time. We map all traceable sessions of these days into a 3-dimensional space of (location, time, volume), and use DBScan to analyze how these sessions are grouped by clustering them regarding the three dimensions. The objective is to understand regular spatial-temporal behaviors of the selected subscribers on making data traffic.

4.1 Mapping Sessions into Three-dimensional Space

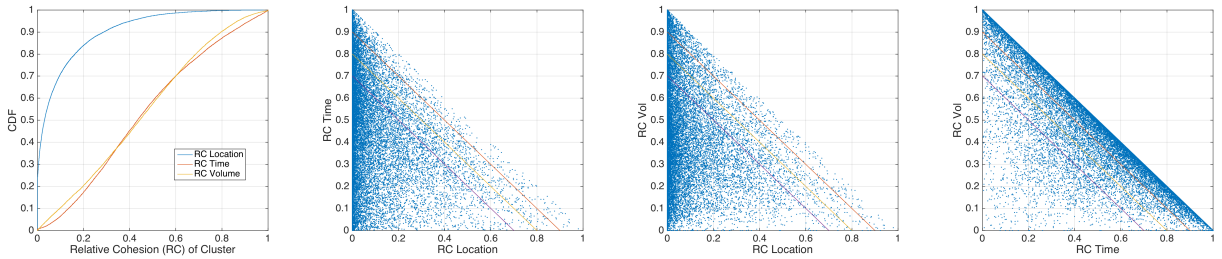
We construct for each subscriber a 3-dimensional space where time, space and volume account respectively for one separate dimension. In the space dimension, each session is represented as its location of cell. Representing a location in coordinates apparently requires two separate dimensions. In order to express a location in one dimension, we cluster all unique cells according to their physical positions using the OPTICS [20] algorithm. OPTICS returns an order of cells, in its reachability-plot. In the order, cells which are spatially closest become neighbors, as Figure 3(a) shows. Hence in the space dimension, a cell is expressed by its rank in the order instead of coordinates. Note that the rank does not affect the further analysis. In the time dimension, a session is expressed as the time of that day. We capture only days from Monday to Friday, since a subscriber may have different behaviors on weekdays and on weekends but have similar behaviors on each weekday [9]. In the volume dimension, a session is represented as its magnitude of volume as $\log_{10} Vol$, in order to shrink the huge range of volume (from 1 to 10^7). For each subscriber, we map all traceable sessions into the 3-dimensional space. See Figure 3(b) as an example. We observe that *a subscriber makes data in a few locations in general, though far more locations may be observed in her mobility.*

4.2 Clustering Sessions using DBScan

Since a subscriber has several weekdays captured, her sessions reflect her regular behavior of generating data traffic on weekdays. We reveal such a behavior by clustering sessions and analyzing how a cluster is aggregated. DBScan [21] is used to discover session clusters. In the DBScan algorithm, we set that a cluster has at least 4 sessions ($minPts = 4$) and $\epsilon = 0.25$ chosen by the k-th nearest neighbor plot (Figure 3(c)) by following [22]. For measuring the distance between two sessions $p_1(\mathbf{l_1}, t_1, v_1)$ and $p_2(\mathbf{l_2}, t_2, v_2)$, an euclidean-like distance is used as $dist(p_1, p_2) = \sqrt{\sum dist^{(*)}(p_1, p_2)^2}$, where the distance in each dimension is separately measured as following: (1) $dist^{(location)}(p_1, p_2) = \omega_l |\mathbf{l_1} - \mathbf{l_2}|_{great-circle distance}$ in kilometers; (2) $dist^{(time)}(p_1, p_2) = \omega_t |t_1 - t_2|$ in hours; (3) $dist^{(volume)}(p_1, p_2) = \omega_v |v_1 - v_2|$ ($|log_{10} \frac{Vol_1}{Vol_2}|$). In order to uniformly measure three dimension uniformly, distances are normalized by the top 1% largest values. The distribution of normalized distances is similar in all three dimensions as Figure 3(d) shows. Most of subscribers has at most 3 clusters (Figure 3(e)) and 3 clusters are enough to cover a majority of sessions (Figure 3(f)), indicating that one may acquire major behaviors of subscriber by analyzing the top 3 biggest clusters.

4.3 Investigating Behaviors from Clusters

The session cluster contains information about the subscriber's behavior. We are interested in the cause of forming a cluster. A cluster can be aggregated due to one or more issues in time, space and volume. Hence we use **relative cohesion** (RC) to quantify the contribution of each dimension on aggregating into a cluster, defined as following: $RC^{(*)} = \frac{\sum_{p \in C} dist^{(*)}(p, \mathbf{c})^2}{\sum_{p \in C} dist(p, \mathbf{c})^2}$, where p is a session and $\mathbf{c}$ is the centralization of cluster. According to the definition of RCs, RC satisfies that $0 < RC^{(*)} < 1$ and $RC^{(loc)} + RC^{(time)} + RC^{(vol)} = 1$. If a cluster has a significantly small RC in one of the dimensions, we can say that this dimension is the reason of aggregating the clusters as *dominating dimension*, *e.g.*, if a subscriber has a cluster aggregated by location, this subscriber is expect to generate data when being observed in a certain range of this location. A cluster has at most two dominating dimensions. As Figure 4(a) shows, most of clusters are aggregated by locations and the rest may be aggregated by volume and time, indicating that *data usage pattern is highly related to subscriber's mobility*, but a subscriber may always generate data on a certain time period, having nothing to do with her



(a) CDF of RC (b) Location versus Time (c) Location versus Volume (d) Time versus Volume

Figure 4: Relative Cohesion (RC) of clusters

whereabouts, since there are time-dominating clusters observed.

In summary, *a subscriber's general behavior of making data traffic is highly related to her mobility.* Especially, *a subscriber generates data traffic only in a few locations.* We will further explore correlation between subscriber's data demand and mobility in future.

5 Summary and Next Steps of the Thesis

We exploit the predictability of data volume of subscriber in the thesis. We observe that the time series of volume contains a significant high degree of potential predictability determined by its information entropy. This high predictability is caused by subscriber's high probability of *idle* and low inherent uncertainty encoded in magnitudes of volume. We also investigate mobility and data demand of subscriber, and observe that subscribers generally generate data traffic in a few locations in spite of large variability in mobility, indicating that mobility has potential to contribute forecasting *idle* and volume.

To the best of our knowledge, there is no work of modeling mobile data or content in terms of both mobility and time. We would like to *not only characterize mobile data but also predict future demand of data volume and content.* Our work will continue in the following directions:

1. **Study spatial-temporal correlations of data volume patterns.** Firstly, we will profile subscriber's volume by following [9], to identify heavy data users. Secondly, we will reveal correlations between subscriber's data usage pattern and location, category subscribers by such correlations, and further link them to the volume prediction.

2. **Extend the predictability analysis to forecast data volume and mobility together.** Since it is observed that mobility is related to data usage pattern. We explore the opportunity of predicting subscriber's data volume and location simultaneously, combing our study and mobility prediction [12, 13].

3. **Analyze predictability of volume based on categorized subscribers.** With the incoming dataset, we are able to distinguish prepaid and postpaid subscribers as well as their operators. We will analyze the differences between the two operators and the two types of subscribers.

4. **Study correlations between content and mobility.** The content represents detailed information of data traffic, such as application type, file type, file size, etc. Understanding subscriber's content demand in data traffic is of great significance for designing caching or prefetching approaches. We will study whether a subscriber has special content demand regarding his locations.

References

[1] C. V. N. Index, "Global mobile data traffic forecast update, 2015-2020," *White Paper, February*, 2016.

[2] E. Mucelli and A. C. Viana, "From routine to network deployment for data offloading in metropolitan areas," in *IEEE SECON 2014*, pp. 126–134, IEEE, 2014.

[3] F. Qian, K. S. Quah, J. Huang, J. Erman, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "Web caching on smartphones: ideal vs. reality," in *ACM Mobisys 2012*, pp. 127–140, ACM, 2012.

[4] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, V. Erramilli, and Y. Grunenberger, "Is there a case for mobile phone content pre-staging?," in *ACM CoNEXT 2013*, pp. 321–326, ACM, 2013.

[5] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: time-dependent pricing for mobile data," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 247–258, 2012.

[6] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *ACM Communications*, vol. 56, no. 1, pp. 74–82, 2013.

[7] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang, "The predictability of cellular networks traffic," in *IEEE ISCIT 2012*, pp. 973–978, IEEE, 2012.

[8] R. Li, Z. Zhao, X. Zhou, J. Palicot, and H. Zhang, "The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice.," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 234–240, 2014.

[9] E. Mucelli Rezende Oliveira, A. Carneiro Viana, K. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," in *IEEE PerCom 2015*, pp. 230–235, IEEE, 2015.

[10] H. H. Jo, M. Karsai, J. Karikoski, and K. Kaski, "Spatiotemporal correlations of handset-based service usages," *EPJ Data Science*, vol. 1, pp. 1–18, 2012.

[11] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale Mobile Traffic Analysis: a Survey," *IEEE Communications Surveys & Tutorials*, vol. PP, no. 99, pp. 1–1, 2015.

[12] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, pp. 1018–1021, Feb. 2010.

[13] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility Supplementary Material." Science Online.

[14] J. Wang, Y. Mao, J. Li, Z. Xiong, and W.-X. Wang, "Predictability of Road Traffic and Congestion in Urban Areas," *PloS one*, vol. 10, no. 4, p. e0121825, 2015.

[15] "Grandata lab, agentina." http://www.grandata.com.

[16] M. Ficek and L. Kencl, "Inter-Call Mobility model: A spatio-temporal refinement of Call Data Records using a Gaussian mixture model.," *IEEE INFOCOM 2012*, pp. 469–477, 2012.

[17] S. Hoteit, S. Secci, S. Sobolevsky, G. Pujolle, and C. Ratti, "Estimating real human trajectories through mobile phone data," in *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, vol. 2, pp. 148–153, IEEE, 2013.

[18] K. S. Kung, K. Greco, S. Sobolevsky, and C. Ratti, "Exploring universal patterns in human home-work commuting from mobile phone data," *PloS one*, vol. 9, no. 6, p. e96180, 2014.

[19] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, *Socio-geography of human mobility: A study using longitudinal mobile phone data*. Computing Science, Newcastle University, 2011.

[20] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," in *ACM Sigmod Record*, vol. 28, pp. 49–60, ACM, 1999.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[22] P.-N. Tan, M. Steinbach, V. Kumar, *et al.*, *Introduction to data mining*, vol. 1. Pearson Addison Wesley Boston, 2006.