Spatiotemporal Individual Mobile Data Traffic Prediction

Guangshuo Chen

Spatiotemporal Individual Mobile Data Traffic Prediction

Guangshuo Chen

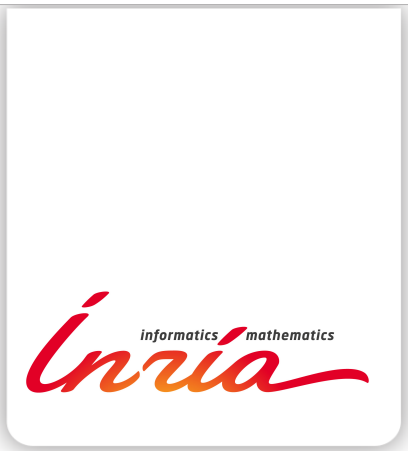
TECHNICAL REPORT

N° 497

Février 2018

Project-Teams INFINE

ISSN 0249-0803 ISRN INRIA/RT--497--FR+ENG



Spatiotemporal Individual Mobile Data Traffic Prediction

Guangshuo Chen*†

Project-Teams INFINE

Technical Report n° 497 — Février 2018 — 21 pages

Abstract: Understanding the nature of data network traffic is critical in network design, management, control, and optimization. In this report, we leverage two large-scale real-world datasets collected by a major mobile carrier in a Latin American country to investigate the prediction of individual mobile data traffic. Based on our previous analysis on the theoretical predictability, we extend our analysis to the actual prediction and validate the findings, that we have observed in the theoretical analysis, in the actual predicting scenario. We implement the typical algorithms for time series prediction in the literature and test their performance. Then, we propose our algorithms based on state-of-the-art machine learning techniques. Our data-driven test on the performance of these predictors shows that a simple Markov predictor can outperform its legacy counterparts in most of the cases. It achieves a mean accuracy of 62%, but it relies heavily on the historical data and can hardly have an enhancement from knowing individual whereabouts. Our proposed solutions can achieve a typical accuracy of 70%, which outperforms all the legacy ones and have a $1\% - 5\%$ degree of improvement by learning individual whereabouts.

Key-words: Cellular networks; mobile data traffic; prediction

This work was supported by the EU FP7 ERANET program under grant CHIST-ERA-2012 MACACO.

* École Polytechnique, Université Paris Saclay, 91128 Palaiseau, France
† INRIA Saclay-Île-de-France, Université Paris Saclay, 91120 Palaiseau, France

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Prédiction spatio-temporelle du trafic de données mobiles individuelles

Résumé : La compréhension de la nature du trafic du réseau de données est essentielle dans la conception, la gestion, le contrôle et l'optimisation du réseau. Dans ce rapport, nous exploitons deux ensembles de données du monde réel à grande échelle collectés par un opérateur de téléphonie mobile majeur dans un pays d'Amérique latine pour étudier la prédiction du trafic de données mobiles individuelles. Sur la base de notre analyse précédente sur la prévisibilité théorique, nous étendons notre analyse à la prédiction réelle et validons les résultats, que nous avons observés dans l'analyse théorique, dans le scénario de prédiction actuel. Nous implémentons les algorithmes types pour la prédiction de séries chronologiques dans la littérature et testons leur performance. Ensuite, nous proposons nos algorithmes basés sur des techniques d'apprentissage automatique de pointe. Notre test basé sur les données sur les performances de ces prédicteurs montre qu'un simple prédicteur de Markov peut surpasser ses homologues traditionnels dans la plupart des cas. Il atteint une précision moyenne de 62%, mais il repose fortement sur les données historiques et peut difficilement être amélioré à partir de la localisation individuelle. Nos solutions proposées peuvent atteindre une précision typique de 70%, ce qui surpasse tous ceux qui existent déjà et ont un degré d'amélioration de 1% à 5% en apprenant les localisations individuelles.

Mots-clés : Réseaux cellulaires; trafic de données mobiles; prédiction

1 Introduction

The understanding of human behaviors is a central question in many research topics and has contributed to a wide range of applications [1, 2, 3, 4, 5, 6, 7, 8, 9]. In cellular networks, human mobility and mobile data traffic consumption are two significant human habits. The ability to understand them has essential implications in many aspects of cellular networks.

- The high availability of mobility prediction can enable various application scenarios such as location-based recommendation, home automation, and location-related data dissemination and also help improving quality of service [6, 8]. In the literature, a large and growing body of literature has investigated the topic of predicting human mobility [1].

- The better understanding of future mobile data traffic demand can help to improve the design of solutions for network load balancing, aiming at improving the quality of Internet-based mobile services [7, 9]. Compared with the human mobility analysis, a far less group of literature has focus on this topic.

In this report, we investigate on the topic of understanding and predicting of mobile data traffic, from the per-user viewpoint. In our last technical report [10], we have analyzed the theoretical predictability of individual mobile data traffic using tools of information theory. In this report, we push our analysis a step further by proposing the design of practical predictors. In particular, We address the problem of understanding spatiotemporal mobile data traffic demand for individuals, and make the following major contributions:

- We implement the major legacy algorithms for anticipating symbolic time series, and evaluate their performance using extensive tests on large scale mobile phone datasets.

- We study on the predictability of per-user mobile data traffic in isolation. Our data-driven results show that practical algorithms that predict from the historical data volumes have high theoretical performance potential, *i.e.*, an expected average prediction accuracy of 81% over the users of study. However, real-world prediction can only achieve up to 65% by the legacy Markovian methods and up to 70% by the machine learning techniques.

- We then extend our study to the predictability with the mobility of each user jointly. We observe that, due to the strong spatiotemporal correlation, forecasting the data traffic and location jointly could achieve a better performance than doing separately. The theoretical analysis reveals that this improvement will be at maximum 10% on average according to [10], Our practical evaluation in this report shows that the machine learning techniques can efficiently leverage the spatiotemporal correlation to improve the prediction accuracy in a degree of $1\% - 10\%$.

- In all, build upon the results in this report, we confirm the findings about the theoretical predictability presented in [10].

2 Mobile Data Traffic Prediction

To what degree is the Internet traffic predictable? It is a question that has led to a number of attractive issues and has been continuously investigated since the invention of the Internet [11]. In this section, we review the state-of-the-art on the prediction of mobile data traffic. Our discussion is organized from two perspectives:

- **Aggregated mobile data traffic.** In this perspective, we consider the mobile data traffic from the viewpoint of a mobile network operator. Such data traffic is aggregated over many mobile devices within the same cell, the same close geographical area, or the same service/application.
- **Individual mobile data traffic.** Here we discuss in an individual viewpoint, *i.e.*, the mobile data traffic that is generated by a single mobile device.

For each perspective, we briefly introduce the data traffic characterization and particularly present the practical prediction techniques. It is worth noting that, in this section, we focus on the studies on the Internet traffic and exclude those on other traffic (*e.g.*, voice calls).

We see that it still lacks the studies of theoretical predictability and actual prediction approaches on the personal view of mobile data network traffic. We have studied the predictability in [10] and focus on the actual prediction in this report.

2.1 Literature on Aggregated Mobile Data Traffic

The investigation on aggregated mobile data traffic is mainly driven by the analyses on worldwide large-scale operator-collected datasets. For instance, such datasets that have nationwide populations are deeply mined in the relevant studies by Paul *et al.* [12] (USA), Hoteit *et al.* [13] (France), and Xu *et al.* [14, 15] (China).

2.1.1 Characterization

There are two major aspects with respect to the characterization, *i.e.*, temporal dynamics and spatiotemporal correlation.

The regularity of the temporal variation of aggregated mobile data traffic is general agreement among the literature [16]. Almost at the same time, Paul *et al.* [12] and Shafiq *et al.* [17] separately investigate the temporal evolution of aggregated mobile data traffic of cell towers and popular applications. They both find that such traffic follows a daily repetitive pattern over weekdays: in general, the traffic has low demand during nighttime and high demand during daytime. The same repetitive pattern is also observed by Xu *et al.* [14, 15]. Is is also remarked in [12, 17, 18] that the traffic over weekdays and weekends have different repetitive patterns and demands; a larger data traffic demand exists on weekdays than weekends. An interesting fact is that the temporal variations observed by Paul *et al.* [12] and Shafiq *et al.* [17] have different peak hours, which is also observed from other network traffic [16]. For this, a possible explanation is that such temporal variation under a higher temporal resolution partially depends on the area of study.

The spatiotemporal correlation exists among the data traffic generated by cell towers over many users in the same area. In the pure spatial perspective, the distribution of the data traffic is spatially heterogeneous: it varies over different regions as revealed by Paul *et al.* [12] and Xu *et al.* [14, 15]. Further, the latter authors find that the cell towers have similar data traffic profiles regarding their regions (*i.e.*, resident, transport, office, and entertainment) and such profiles of adjacent cell towers are correlated. In the spatiotemporal perspective, the two papers show that the spatial heterogeneity above also varies over time: the pick hours depend on the regions. The former authors leverage a quantitive measure (*i.e.*, the Moran's I statistic) to evaluate spatiotemporal diversity of the data traffic. They find that in general, the imminent loads of adjacent cell towers are more correlated when these loads are high, but the correlation is relatively weak and almost disappears around midnights. Recently, [19] further investigates the spatiotemporal correlation and propose an approach to infer the hidden spatial and temporal structures of aggregated mobile data traffic.

Also, several studies reveal the spatial heterogeneity aggregated over applications. The earlier work by Trestian *et al.* [20] already shows that the Internet traffic over services and applications is consumed differently at home and work locations. Hoteit *et al.* [13] find that the data traffic loads of cell towers have different inner diversities among TCP- and UDP-based services. Later, the extended analysis by Shafiq *et al.* [21] finds that the data traffic aggregated by popular applications is strongly heterogenous over regions. This provides the capability of categorizing cell towers into four classes (web browsing, email, audio, and mixed traffic) with respect to the major applications in their data traffic loads.

2.1.2 Prediction

Some efforts have been put on the prediction of aggregated mobile data traffic. They aim at converting the observed dynamics and correlations above to practical prediction techniques. In the following, we review the proposed prediction techniques according to the level of the aggregation.

- **Cell-level data traffic**. There is a common observation on the fact that the data traffic of cell towers has a high degree of both theoretical and practical predictabilities. Regarding the theoretical viewpoint, Zhang *et al.* [22, 23] investigate the limits of the theoretical predictability by observing the traffic of $7,000$ cell towers in China. They find that under the temporal resolution of 30 minutes, aggregated traffic (voice, text, and data) can be well predicted from the historical demand of the preceding 15 hours; the theoretical predictability of the data traffic is lower than that of the data flow of voice calls or text messages. They also find that the knowledge of the traffic demands of adjacent cells towers can enhance the theoretical predictability, but in a less degree on the data traffic than the others, which supports the quantitative evaluation on the spatiotemporal correlation by Paul *et al.* [12]. Their results ensure the capability of time series prediction techniques on the prediction of such traffic.

  Regarding the practical prediction techniques, Xu *et al.* [14, 15] show that the cell-level data traffic is predictable via a linear combination of four primary components corresponding to human activities. Zang *et al.* [24] propose a mixed machine learning approach composed of K-means clustering, Elman Neural Network, and wavelet decomposition. An alternative prediction approach is proposed by Yi *et al.* [25]; it builds a complex network for all the cell towers, measure the traffic on the very important ones, and predict the others' traffic using Support Vector Regression – another machine learning method. It can recover the whole picture of the traffic demand from only 8% of the total cell towers. In the opposite viewpoint, Nika *et al.* [26] perform an empirical study on data hotspots using a large-scale operator-collected dataset of $5,327$ cell towers, and show the availability of standard machine learning methods on the prediction of future hotspots (cells towers) of the traffic demand from the past history.

- **Application-level data traffic**. The early paper by Keralapura *et al.*[27] proposes a technique to cluster users and their browsing profiles. The authors find that user behavior in terms of Internet surfing can be captured using a small number of clusters. Such heterogeneity of aggregated mobile data traffic is also explored by Ying *et al.* [28]. Later, Shafiq *et al.* [17] uses a Zipf-like model to capture the distribution of application-level mobile data traffic and finds that the regularity makes the temporal variation of the traffic highly predictable from the history of the past demand using a simple Markovian method. Recently, Zhang *et al.* [29] design a mixed application-level traffic prediction framework that leverages the α-stable modeled property and dictionary learning to separately deal with

the temporal variation and the spatial sparsity of the traffic. Marquez *et al.* [30] extend the analysis in [17] and reveal a strong heterogeneity in difference mobile service demands using correlation and clustering. They show that the temporal usage patterns are quite different from service to service. Besides, several works focus on the traffic generated by special services, such as chatting (*e.g.*, WhatsApp [31] and WeChat [32]), video streaming [33], and mobile cloud [34].

In summary, the proposed techniques extend the technical bound on the prediction of mobile data traffic: they not only leverage the legacy tools that used for analyzing wired network traffic (*e.g.*, the entropy, Markov property, α-stable modeled property) to capture the temporal variation but also import several state-of-the-art machine learning tools to utilize the spatiotemporal correlation.

2.2 Literature on Individual Mobile Data Traffic

A relatively small body of literature is on the investigation of individual mobile data traffic, which is also driven by the data mining. Differently, the relevant studies utilize both large-scale operator-collected datasets, *e.g.*, by Paul *et al.* [12] and Oliveira *et al.* [18, 35], and small-scale mobile crowdsensing datasets, *e.g.*, by Jo *et al.* [36].

2.2.1 Characterization

The characterization from the individual viewpoint is performed by Paul *et al.* [12], Jo *et al.* [36], Li *et al.* [37], Oliveira *et al.* [18, 35], among others.

There is an general agreement on the heterogeneity of the data traffic, with respect to the user population and the time. It is shared by Paul *et al.* [12] and Oliveira *et al.* [18, 35]. They show that most of the total data traffic is generated from a small group of "heavy" users.

Regarding the temporal variation, both the authors above find that, in general, each user is highly active only in a few hours per day, and similarly, the temporal variation is different on weekdays and weekends, as in aggregate mobile data traffic. The latter authors [18, 35] find that individual mobile data traffic also follows daily repetitive patterns and the users also have peak and non-peak hours in terms of the data traffic. In particular, they find that the variation of different hours within the same day is stronger than that of the same hours overs different days.

As to the spatiotemporal correlation, Paul *et al.* [12] point out that a user is usually active at only a few of his common locations. Jo *et al.* [36] mine a small dataset of locations and services of 124 users over 16 months and they identify the spatiotemporal correlations of service usage patterns.

Other dynamics with respect to social features are also revealed. For instance, Oliveira *et al.* [18, 35] find that the distribution of individual mobile data traffic is slightly heterogeneous over the age and gender; Li *et al.* [37] focus on the major smartphone operating systems and discuss the traffic dynamics and major application in each system.

2.2.2 Prediction

Yet, fewer studies have addressed the prediction of individual mobile data traffic. Regarding the bandwidth of mobile devices around a cell tower, a theoretical analysis is performed by Bui *et al.* [38, 39]. Based on a theoretical LTE radio model, the authors propose a model to predict the bandwidth of mobile devices over a wide range of time scales [38]. Their model considers both the user location and the statistic of bandwidth availability. They also design a refined model aiming at the prediction of short-term bandwidth using Gaussian Random Walks [39].

Regarding the latency of each data session, Zhao *et al.* [40, 41] address the static and dynamic latency estimation problems and propose a distance-feature decomposition algorithm based on the Matrix Factorization technique to predict the latency.

In summary, the current literature has already shown the temporal dynamics and spatiotemporal correlations of both aggregated and individual mobile data traffic, while only several practical prediction techniques are proposed aiming at the latter's prediction. In this context, a large amount of effort has to be put on the literature. For instance, to perform data-driven prediction analysis on the theoretical and practical predictabilities of individual mobile data traffic.

3 Prediction Methods

4 Data preliminary

4.1 Datasets

Our study is based on two real-world datasets describing the cellular network activity of hundreds of thousands of mobile phone subscribers (identically called users) of a major cellular operator in a metropolitan area. All data refers to a consecutive period of 1 year. The first dataset consists of *call detail records (CDRs)* containing timestamped and geo-referenced logs (*i.e.*, of the closest mobile cell tower) of each voice call performed by every user. The second dataset describes the *Internet data sessions* established every time a mobile device needs to exchange IP data traffic through the cellular network.

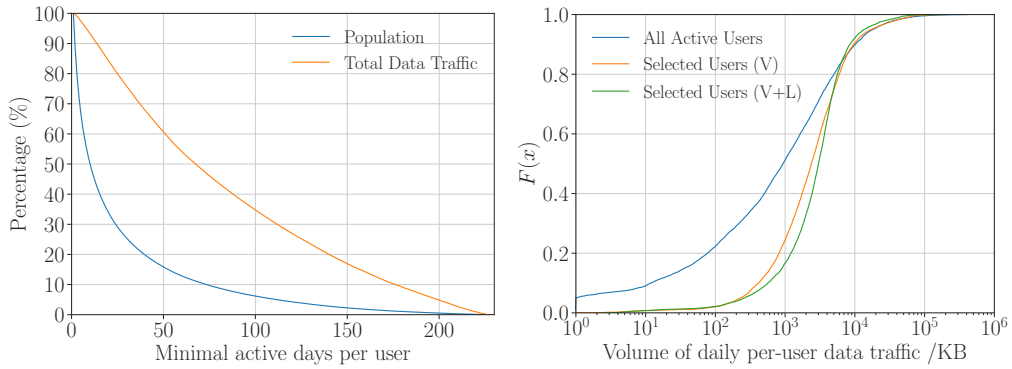
These two datasets provide different and complementary information: CDR data includes location information that allows reconstructing user mobility, while session data only presents the mobile data traffic volume generated by each subscriber (with no associated geo-referenced log). In both cases, we preprocess the datasets to construct time series of subscriber's locations and data traffic demands that are representative and statistically significant.

- `CDR` dataset. Call detail records are logged every time a mobile device makes or receives a voice call. Each entry contains the hashed identifiers of the caller and callee, the call duration in seconds, the timestamp of the call start time and the location (latitude and longitude) of the cell tower to which the device is connected when initiating the phone call.

- `Session` dataset. Every Internet data session is established upon the allocation of a radio channel for the exchange of IP traffic, and it ends after an idle period over the same channel. Each entry in the dataset contains the hashed device identifier. The same hashing function is used in the CDR and Internet data session datasets, which allows linking users in the two datasets. The volume of upload and download data exchanged in KiloBytes, and the timestamp denoting the starting time of the session. The dataset does not contain spatial information.

4.2 Per-user Spatiotemporal Data Construction

We rely on two datasets (*i.e.*, `CDR` and `Session`) for the mobility and mobile data traffic information respectively. There is no readily choice of appropriate users or time series. We have to extract them via a data-driven analysis.



(a) (b)

Figure 1: (a) Distributions of the number of minimum active days versus the number of users and the ratio of mobile data traffic. (b) CDF of the mean daily mobile data traffic of each user over all the users and the selected users.

4.2.1 Active User Selection

Regarding the user selection, the basic rationale is as follows. First, we have to focus on the users who actively generate Interest data traffic through their mobile devices. Second, we need to reveal the full performance of utilized prediction methods under normal circumstances. For that, each user needs to have enough "regular" days in which he generates mobile data traffic. Therefore, we focus on weekdays and exclude holidays, vacation periods, and weekends (having insufficient days and different data traffic patterns). Third, we need to study the predictability along with mobility. Thus, each selected user must have enough location samples (*i.e.*, voice call CDR in the `CDR` dataset) to build his slotted CDR-based trajectory. In the following, we analyze the `CDR` and `Session` datasets and proceed the user selection.

We first discuss the criterion of the user selection with respect to mobile data traffic. During the observing period shared by the two datasets, there are 229 regular weekdays in total. We consider the daily activeness of each user in these weekdays. For that, an *active* day is defined as a day in which a user generates at least 1KB of mobile data traffic. We then portray in Figure 1(a) (blue line) the minimum number of active days versus the percentage of users. We observe that nearly 85% of the users are "inactive": they only generate data traffic in less than 50 days despite a 15-month observing period. For comparison, we plot in the same figure (orange line) the ratio of the total data traffic generated by these users. We see that these "inactive" users only take account for 40% of the total mobile data traffic, while approximately 5% of the users who have at least 150 active days generate almost 20% of the total mobile data traffic. The observation above confirm the existence of the so-called "heavy" users, as in [12, 18]. In our case, we need to focus on the heavy users and to have their time series of data traffic volumes as long as possible. Considering both the user activeness and the available number of users, we choose the users who have more than 150 active days, which provides us approximately 92K of the selected users.

We validate the criterion above in terms of the per-user data traffic consumption. Figure 1(b) portrays the CDF of the mean daily data volume of each user. We see that the distribution of our selected users (orange line) implies a more positive usage of mobile data traffic than that of all the users: 80% of the former users generate at least 1MB per day on average, while only 50% of the latter users do the same. Consequently, these selected users are "heavy" users which we

need; they are highly active every weekday and generate a large amount of mobile data traffic.

Second, we have to select users in terms of the number of locations of each user. We need to build complete slotted CDR-based trajectories, in order to perform the joint predictability analysis. Due to our analysis on the completeness of the `CDR` dataset, there is no readily CDR-based trajectory having 100% of the completeness. Therefore, our slotted CDR completion technique proposed in [42] appears as a rescue, which is later used to reconstruct full complete trajectories. Our technique achieves its best average performance when the trajectories to be completed have at least 20% of the completeness. Besides, the completeness analysis shows that the number of available users having at least 20% of the completeness of is extremely small under high temporal resolutions. Thus we choose the temporal resolution of one hour. In summary, our user selection criterion with respect to mobility information is that a user should have a slotted CDR-based trajectory having at least 20% of the completeness. Given this criteria, 7K users are selected from the overall 92K "heavy" users. These users, as shown in Figure 1(b), have the distribution of daily mobile data traffic volumes (green line) highly close to that of all the "heavy" users. It indicates that adding the mobility criteria into the user selection does not impact the activeness of mobile data traffic consumption heavily in the selected users.

4.3 Discretization of Volumes and Locations

Regarding mobile data traffic, for each select user u, we construct a time series represented by $v_1^T(u) = \{v_1^u, \cdots, v_T^u\}$, where v_i^u is his discrete mobile data traffic generated during the i-th time slot. We consider the temporal resolution of one hour. It means that here are 24 time slots on each day and mobile data traffic is computed on an hourly basis: This will be our default setting. For the discretization of data traffic volumes, we favor a representation that captures the data traffic magnitude over a uniform discretization. The rationale is that one is more interested in predicting whether a user will generate, *i.e.*, KiloBytes, MegaBytes or GigaBytes of traffic, rather than if a user's demand will be in the first (1 KB, 333 MB), second (334 MB, 666 MB) or third (667 MB, 1 GB) portions of one GB. Specifically, we employ the quantization of the data traffic volume spectrum as follows: Eight quantization levels, *i.e.*, 0, $(1, 10)$, $(10, 10^2)$, ..., $(10^6, 10^7)$, all values in KB.

Regarding mobility, each selected user u is collected as a time series of discrete locations, represented by $\ell_1^T(u) = \{\ell_1^u, \cdots, \ell_T^u\}$, where ℓ_i^u is the user's *representative location* of the i-th time slot. Such a time series can be converted from the time series of CDRs with their corresponding cell tower identifiers. Note that the distance between each two discrete locations is still measurable as we have their geographical coordinates. In particular, each day is split into 24 time slots as our default temporal resolution of data traffic volumes; each representative location is selected on an hourly basis. Even then, there is no readily full complete CDR-based trajectory that can be extracted from the `CDR` dataset. For that, we apply our proposed slotted CDR completion technique on the incomplete CDR-based trajectories of the selected users, and then convert them to time series of discrete locations.

In summary, we have two criteria of the user selection corresponding to mobile data traffic and mobility, respectively. Two groups of the "heavy" mobile data traffic users are then selected given the criteria, as shown in Table 1. The first user set $\mathcal{U}_1$ contains 92K users and their data session records extracted from the `session` dataset. For each user $u \in \mathcal{U}_1$, we have his time series of discrete mobile data traffic. The user set $\mathcal{U}_1$ and its data will be used in the predictability analysis using temporal dynamics in Section 6. The second one $\mathcal{U}_2 \subset \mathcal{U}_1$ consists of 7K users and has their locations extracted from the `CDR` datasets in addition. For every user $u \in \mathcal{U}_2$, we have his time series of locations and quantized data traffic volumes in the temporal resolution of one hour. This data will appear in Section 7.

Table 1: Users of Study

Group	Population	Time Series	Resolution	Days
User set $\mathcal{U}_1$	92K	$v_1^T(u)$	$\{15, 30, 45, 60\}$Min	≥ 150
User set $\mathcal{U}_2$	7K	$v_1^T(u)$, $l_1^T(u)$	60Min	

5 Prediction Methods

5.1 Markovian Prediction Methods

The methods in this class, *i.e.*, MC, PPM, SPM, and ALZ, leverage the Markov property. They are mainly designed for the prediction of time series of discrete observations. Their application assumes that the target time series has the Markov property, *i.e.*, its current value is always determined by a limited number of its previous values. In this class, a prediction method predicts the current value X_t of a time series by building a probabilistic model from its full history and solving the following maximization problem: $\hat{x}_t = \arg\max_x P(X_t = x | x_{t-1}, \cdots, x_{t-k})$ where $x_{t-1}, \cdots, x_{t-k}$ are the newest k previous values observed in the time series.

MC (Markov Chain) This is almost the simplest Markovian method [43]. A k-th order Markov chain, represented as MC(k), makes a prediction of the state X_t solely based on the fixed previous k states. It builds a transition matrix consisting of the probabilities of transitions from the past k states to the current one. There are several common practices to compute the probabilities, such as MLE (maximum likelihood estimation) [44] and MCMC (Markov Chain Monte Charlo) [45]. However, it needs a large number of samples to compute probabilities, which grows quickly with respect to the order k and the alphabet of discrete values.

PPM (Prediction by Partial Matching) This is an improved method of the Markov chain, used massively in the lossless text compression [46]. A k-th order PPM model, represented as PPM(k), is a combination of MC(m), $\forall m \leq k$. It computes the so-called *escape* probabilities of the state X_t as the weighted sums of the probabilities of all the Markov models in the combination. An example of the PPM model is given as follows. Note that we use the implementation of Moffat *et al.* [46, Method C] in this thesis.

SPM (Sampled Pattern Matching) This is another improved method of the Markov chain designed by Jacquet *et al.* [47]; for predicting the current state X_t, it considers much larger immediately preceding states than the MC or PPM does. In a SPM predictor, instead of using a fixed order k, the considered length of immediately preceding context is determined as a fixed fraction (represented as the parameter α) of the longest context which has previously appeared. A SPM model with the parameter α is represented as SPM(α).

ALZ (Active LeZi) This is an improved online prediction algorithm based on the classical LZ78 data compression scheme proposed by Gopalratnam *et al.* [48]. It also employs the power of the Markov property and is able to incrementally learning the sequence and to deliver real time predictions. The ALZ algorithm also makes a prediction of the state X_t in the time series given the preceding context, while a variable window of immediately preceding symbols is maintained, of which the length is the longest phrase previously observed in a classical LZ78 parsing. With this window, the algorithm can compute statistics on all possible preceding contexts. For the pseudo code of this algorithm, we refer the reader to [48, Figure 3].

5.2 Machine Learning Methods

This class contains a group of state-of-the-art techniques that are categorized to the field of *supervised machine learning* in practice [49]. These techniques can solve problems in the shape of $\mathbf{y} = f(\mathbf{x})$ where $\mathbf{x}$ and $\mathbf{y}$ are the input and output vectors. Each of them builds a model (which is composed of kernel functions, decision/regression tress, or neuron networks) upon a training set that consists of known instances of $\mathbf{x}$ and its corresponding $\mathbf{y}$ as the output classes (in a classification problem) or values (a regression problem). Then, the trained model can predict $\mathbf{y}$ from $\mathbf{x}$ in a new instance. They are capable of forecasting time series of continuous and discrete values.

MLP (Multilayer Perceptron) This algorithm is a typical supervised learning algorithm using artificial neural networks. It is designed for both regression and classification problems. A MLP network is a feedforward artificial neural network that is fully connected. The MLP algorithm accepts different activation functions, layers and neurons [50].

6 Investigation through Temporal Dynamics

We study the predictability of mobile data traffic generated by individual users. For now, we focus on the forecasting scenario of data traffic volumes in isolation Note that we have already evaluated the theoretical predictability, *i.e.*, the maximum accuracy that any algorithm has potential to achieve in the prediction of individual mobile data volumes [10], on the same group of users as introduced in Section 4. As necessary knowledge, we recall the major findings of the predictability analysis as presented in [10]:

- We find that, by just considering temporal correlations in the traffic, 81% of the activity of each user can be anticipated on average. We prove the result above to hold across heterogeneous classes of subscribers, based on age, gender, mobility, or mobile service usage.

In this report, we put the theoretical results above into the practical performance. In the following, we address whether or not the high theoretical predictability can be achieved. We evaluate the practical predictability of several predictors in the real-world prediction of mobile data traffic volumes generated by the users of the set $\mathcal{U}_1$.

6.1 Methodology

We compute the practical predictability. We rely on actual prediction methods (or in short, *predictors*) to forecast human behaviors. Although the predictability of a human behavior is determined by its uncertainty in substance, it is shown through the performance of predictors on the surface. Therefore, we define the practical predictability with respect to each predictor.

Theoretical predictability Given a human behavior represented as finite discrete values, its practical predictability $\pi^{predictor}$ that corresponds to a real-world predictor is defined as the probability that this predictor can correctly forecasting the behavior's current value. In our setting, given a human behavior of a user u, we have its T observations as a finite time series $x_1^T \equiv \{x_1, \cdots, x_T\}$. Suppose that a predictor uses the first T_s values (*i.e.*, $\{x_1, \cdots, x_{T_s}\}$) to initialize itself and makes predictions of the remaining time slots. We estimate the practical

predictability $\pi_u^{predictor}$ as the observed prediction accuracy using the $\{x_{Ts+1}, \cdots, x_T\}$ values as ground-truth, *i.e.*, mathematically,

$$\pi_u^{predictor} = \frac{1}{|T - T_s|} \sum_{t=T_s+1}^{T} \mathbb{1}(x_t = \hat{x}_t | x_1^{t-1}), \tag{1}$$

where x_t and $\hat{x}_t$ are the actual and predicted values in the t-th time slot.

For each predictor and each user's time series, we initialize the predictor using data in a number of the time series' beginning days and evaluate the average prediction accuracy (defined in Equation (1)) on the rest of the time series as our estimation of the practical predictability. Recall that each user has a time series of discrete data traffic volumes collected in at least 150 days. To let each predictor "warm up" entirely and to exclude the accuracy impact brought by the lack of enough samples, we employ the first 100 days (*i.e.*, $T_s = 100 \times 24$ in Equation (1)) in the initialization of all the predictors. To ensure prediction substantial accuracy, we update each predictor periodically in the evaluation. In particular, given a user's time series of data traffic volumes $v_1^T(u)$, the practical predictability of a predictor is computed by the following procedure:

(1) Initialize the predictor using data from the first 100 days, *i.e.*, the partial time series $v_1^{T_s}(u)$ and then set $D = 100$ days.

(2) Use the predictor to make predictions of all time slots in the $(D+1)$-th day.

(3) Update the predictor using the data traffic volumes generated in the $(D+1)$-th day and set $D = D + 1$.

(4) Go back to (2) if D does not exceed the last day of $v_1^T(u)$. If it exceeds, stop the iteration and compute the practical predictability $\pi(\mathcal{V})$ defined according to Equation (1) .

In this section, we employ this procedure on the time series of data traffic volumes and the predictors introduced later on. Note that this procedure also holds for the remaining practical predictability analyses in this report, *i.e.*, to be applied on the time series of both data traffic volumes and locations.

6.2 Predictors

Building upon the procedure above, we evaluate the practical predictability $\pi(\mathcal{V})$ of several predictors. Since the theoretical predictability upper bound shows the highest expected performance of any predictor that leverages the regularity hidden in the temporal orders of a time series, we mainly choose our predictors that utilize such regularity, which are listed as follows.

- **Markovian predictors**. We utilize all the Markovian predictors presented in Section 5.1, *i.e.*, the `PPM`, `MC`, `SPM`, and `ALZ` predictors. The `PPM` and `MC` predictors make a prediction of the current data traffic volume from the preceding k previous data traffic volumes, the `SPM` predicts from the ratio α of the longest preceding data traffic volumes that appears previously, and the `ALZ` decides the length of the preceding data traffic volumes via an automatic sliding window. Following the previous experience of the application of `PPM`, `MC`, and `SPM` on predicting locations [44] and aggregated data traffic volumes [17], we set their corresponding parameters as follows. For `PPM`(k) and `MC`(k), we choose $k \in [1, 5]$; for the `SPM`(α), we choose $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Regarding their implementations, the `MC` builds a transition matrix of probabilities by employing simple Markov chains and

maximum likelihood probability estimation. For the `PPM` and `SPM` predictors, we favor their implementations in [46, Method C] and [47], respectively. The `ALZ` follows the algorithm presented in [48, Figure 3].

- **MLP (Multilayer Perceptron)**. This is the most classical machine learning technique that leverages artificial neural networks [50]. It has well-tested implementation and good flexibility, which can be even deployed on mobile devices equipped with mobile AI hardware. In particular, we employ a fully connected neural network that has three hidden layers, where each layer has 256 neurons and is activated by the ReLu function [50]. In the training phrase, the network is trained by the Adam optimizing method [51] with the initial learning rate 0.001. Regarding network input and output, we distinguish two predictors based on the MLP.
  - The `MLP` predictor has the same input and output format as in the `PPM` or `MC` predictor, *i.e.*, the preceding k data traffic volumes as input and the prediction of current data traffic volume as output. Compared with the Markovian predictors, it can accept a larger k and thus, we set $k \in [1, 8]$.
  - The `MLP-CI` predictor further employs the temporal contextual information as input along with the preceding data traffic volumes. Particularly, its input vector consists of k daily vectors that represent the data traffic consumption of a user in the previous days. Similarly, we set $k \in [1, 8]$. Each daily vector contains the discrete data traffic volume, the weekday via one-hot encoding, the time slot's hour, and the time difference with the target time slot in hours and days respectively. This predictor still generates a prediction of the current data traffic volume as output.

6.3 Prediction Results

Our results with respect to the practical predictability $\pi(\mathcal{V})$ of discrete data traffic volumes of each user $u \in \mathcal{U}_1$ are shown in Figure 2. For the per-user prediction accuracy of the `PPM`, `MC`, `SPM`, `MLP`, and `MLP-CI` predictors regarding their possible parameters, we plot the CDF of the prediction accuracy of each user categorized by the different settings of the same predictor in Figure 2(a-e). We observe that the performance of these predictors varies slightly with different settings. Particularly, the `PPM` and `MC` achieve their overall best performance when $k = 2$, so does the `SPM` when $\alpha = 0.25$. The reason is that the Markovian predictors have large probability space that increases quickly following a power law with the order k. Therefore, when $k > 2$, these predictors may suffer from lack of sufficient samples. Correspondingly, the `MLP` and `MLP-CI` achieve their best when $k = 4$, indicating the advantage of machine learning techniques clearly, *i.e.*, they can accept larger preceding data effectively in the prediction. Overall, we see that on each particular prediction, importing more historical data does not significantly enhance the prediction accuracy.

In our case, all the predictors are applied on a per-user basis, which means each user may have different setting of a predictor to have his own best prediction accuracy. For that, we employ a 3-fold cross validation process during the initialization of each predictor, to determine the best setting of each user. Here the practical predictability $\pi(\mathcal{V})$ of a certain predictor represents the best performance that it is achieved by each user on his own setting. By merging the results above, we portray in Figure 2(f) the CDF of the practical predictability $\pi(\mathcal{V})$ of each predictor in the prediction of discrete data traffic volumes, where we observe the following:

- Even the worst predictor, *i.e.*, `ALZ`, can still achieve the average prediction accuracy of 55%, which is approximately 10% below the best predictor and 26% below the theoretical

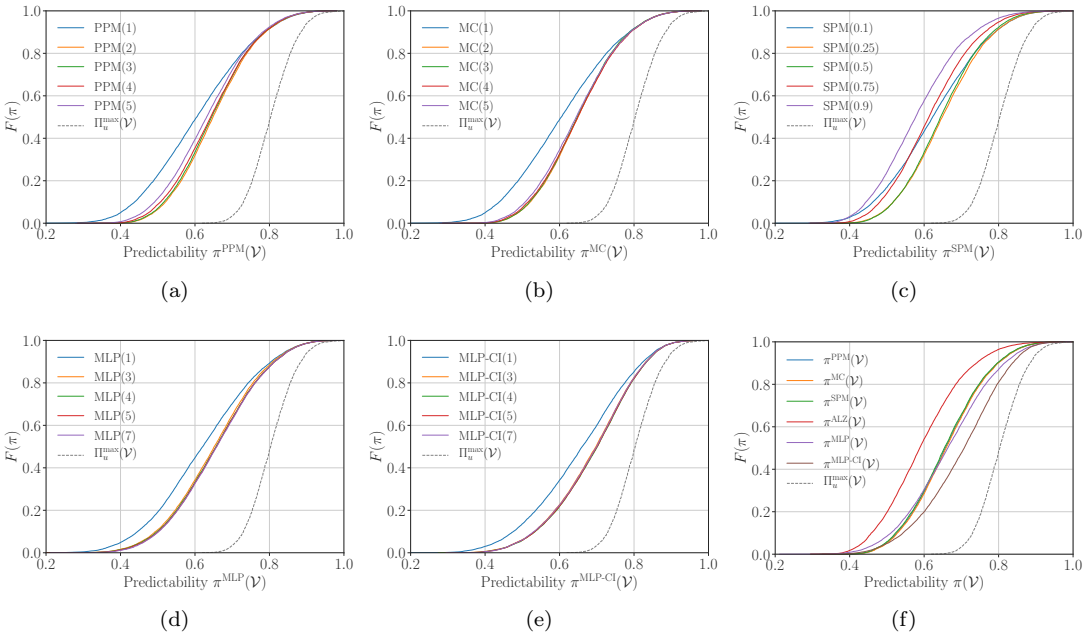


Figure 2: (a-e) Distributions of the prediction accuracy of each predictor with respect to its parameters. (f) Distribution of the practical predictability $\pi(\mathcal{V})$ of each predictor across the user set $\mathcal{U}_2$.

predictability upper bound. This is consistent with the theoretical lower bound $\Pi_u^{\min}(\mathcal{V})$ (*i.e.*, the average accuracy of 42% in theory).

- The other Markovian predictors (`PPM`, `MC`, and `SPM`) have almost the same distributions of the prediction accuracy. This is reasonable because they are all designed upon the Markov chian. Particularly, they all have the mean prediction accuracy of 65%, which is still 16% below the upper bound .
- The `MLP` performs slightly better than the Markovian ones, achieving the average accuracy of 67%. In the distribution, the prediction accuracy per user varies more heavily than the latter. Combining the results, we conclude that among the predictors that only employ the regularity of the temporal orders of discrete data traffic volumes, it is hard to select one which has noticeable advantage over the others. In this context, although the `MLP` performs better, the simple `MC` is quite sufficient having a good trade-off between the computing complexity and achieved performance.
- The overall best performance comes from the `MLP-CI` predictor, which achieves the average prediction accuracy of 70%. Compared with the others, this predictor uses the temporal context as input, which provides more information in each time slot and capture the temporal regularity of mobile data traffic in a better manner.

Consequently, our results confirm that the high degree of the theoretical predictability is consistent with the practical predictability. Even a simple Markovian predictor can achieve a fairly good performance in the prediction of per-user mobile data traffic, which is consistent with the observation on the aggregated mobile data traffic [17]. The machine learning technique

has potential to further improve the prediction performance, while additional information is necessary to have a leap on the prediction performance.

7 Investigation through Spatiotemporal Dynamics

In this section, we push our analysis further to the study of the joint predictability of mobile data traffic volumes and visited locations on a per-user basis. We investigate how predictable is the combination of *how much* traffic is generated by a mobile phone user and *where* this happens. Our analysis provides a comprehensive understanding of whether it is possible to anticipate when, where, and how much mobile data traffic is generated by individual users. It is worth noting that, in this section, our analysis is based on the user set $\mathcal{U}_2$.

We check whether the theoretical predictability (presented in [10]) is consistent with the practical performance. Following the methodology presented in Section 6.1, we evaluate the practical predictability of forecasting each "current" data traffic volume and visited location jointly.

7.1 Excluding Less Frequent Locations

For this analysis, we preprocess the time series of locations. For each user, we keep the most frequent fifteen locations in his time series and merge the rest into one "fake" location marked as "other." Thus, each time series $\ell_1^T(u)$ has at most 16 unique locations. This is to reduce the size of the probability space which increases with the order k and the number of unique locations in our predictors (*e.g.*, an MC(k) predictor needs to build $(N_\ell * N_v)^{k+1}$ probabilities in its transition matrix). We choose the threshold $k = 15$ due to our observation in Figure 3(a), where we see that the top 15 locations can occupy 95% of the time slots in the observing period. In this and next sections, we always employ this top-15 version of the time series $\ell_1^T(u)$ of our users instead of the original ones.

7.2 Prediction Results

Our evaluation on the joint practical predictability is still based on the predictors previously used and presented in Section 6.2, *i.e.*, the `PPM`, `MC`, `SPM`, `ALZ`, `MLP`, and `MLP-CI` predictors. Recall that we have for each user $u \in \mathcal{U}_2$ a mixed time series $m_1^T(u)$ consisting of $v_1^T(u)$ and $\ell_1^T(u)$. Based on these mixed time series, we proceed as follows.

First, we perform the procedure presented in Section 6.1: we initialize our predictors by the partial mixed time series $m_1^T(u)$ of the first 100 days and then predict the (volume,location) pairs in the rest of the mixed time series. In this case, the joint practical predictability $\pi(\mathcal{V}, \mathcal{L})$ that we compute matches to the same joint forecasting scenario as the joint theoretical predictability $\Pi_u^{\max}(\mathcal{V}, \mathcal{L})$ above. Particularly, for the mixed time series, a success prediction of a time slot has to be correct in both the data traffic volume and visited location of that time slot. Our results with respect to the joint practical predictability $\pi(\mathcal{V}, \mathcal{L})$ are shown in Figure 3(b), which we observe the following.

- Among the predictors that only leverage the historical temporal orders of the mixed time series of each user, the `MLP` predictor performs the best with a quite little advantage over the others, while the `ALZ` does the worst. The other three Markovian predictors have almost the same performance.
- The `MLP-CI` predictor achieves the overall highest joint practical predictability; it has 50% of the average prediction accuracy to correctly forecast the data traffic volume and location

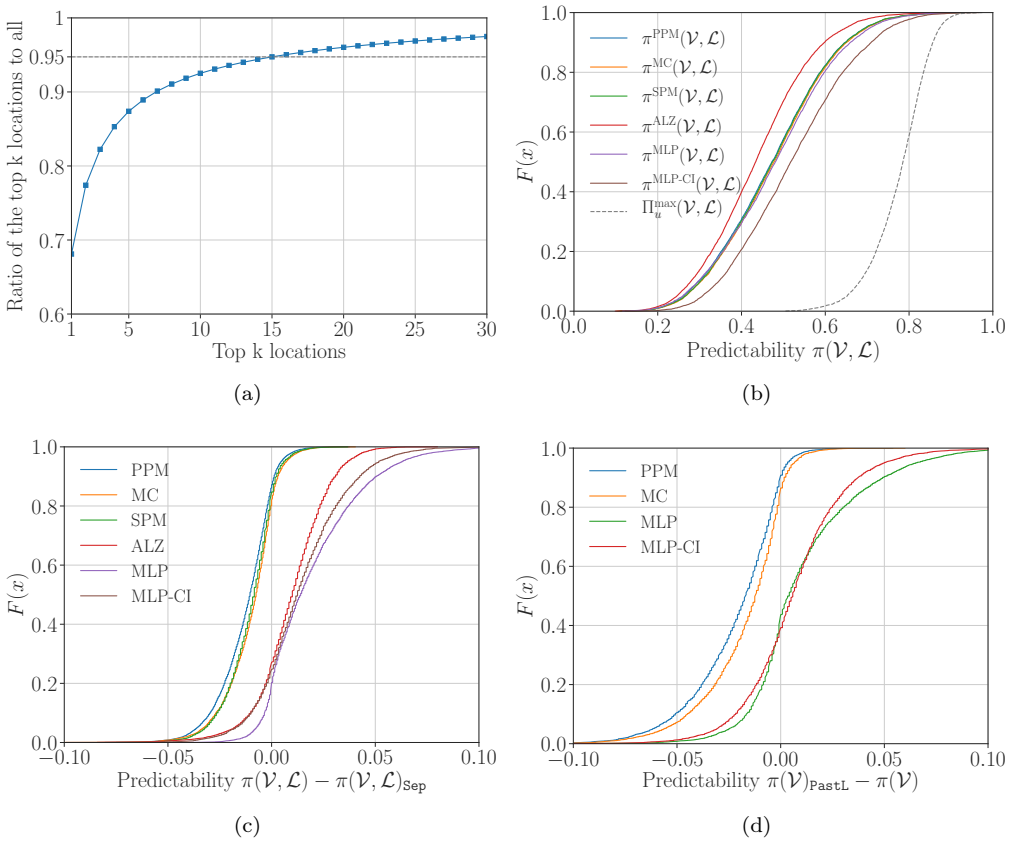


Figure 3: (a) Ratio of the top k locations to all the locations in the time series of locations owned by the user set $\mathcal{U}_2$. (b-d) Distributions of (b) the joint practical predictability $\pi(\mathcal{V},\mathcal{L})$ across the user set $\mathcal{U}_2$, (c) the enhancement on the practical predictability $\pi(\mathcal{V},\mathcal{L})$ by forecasting data traffic volumes and visited locations jointly compared with doing separately, and (d) the enhancement on the practical predictability $\pi(\mathcal{V})$ of each predictor by adding the information of the historical visited locations as a prior knowledge.

at each time slot simultaneously. As in the forecast of data traffic volumes in isolation, the improvement of the `MLP-CI` compared with `MLP` comes from the temporal contextual information.

- Still, there is a larger gap between the theoretical and practical predictabilities in the joint forecasting scenario than only predicting data traffic volumes. Even the performance of the best predictor is still far (*i.e.*, 30% on average) from the theoretical upper bound.

For comparison, we also evaluate the joint practical predictability *with the separate forecasting scenario*, represented by $\pi(\mathcal{V},\mathcal{L})_{\texttt{Sep}}$. For that, we employ our predictors to predict the data traffic volumes and locations separately. Then we combine the predicted volume and location of each time slot into a prediction of the mixed time series. In this case, the maximum value of $\pi(\mathcal{V},\mathcal{L})_{\texttt{Sep}}$ is limited by $\Pi_u^{\max}(\mathcal{V}) * \Pi_u^{\max}(\mathcal{L})$ in the theoretical predictability analysis. Compared with the joint forecasting scenario, we show the improvement of each predictor by computing $\pi(\mathcal{V},\mathcal{L}) - \pi(\mathcal{V},\mathcal{L})_{\texttt{Sep}}$ in Figure 3(c), from which we can clearly see the following.

- Regarding the Markovian predictors, only the `ALZ` can benefit from predicting the two behaviors jointly: almost 80% of the users have their improved performance up to 8%, while the problem is that this predictor performs the worst compared with the others. For the rest three Markovian predictors, only 20% of the users have better performance up to 3% of the improvement. A possible reason for this is that the number of samples is insufficient as the joint forecast scenario enlarges the probability space significantly.

- The `MLP` and `MLP-CI` predictors are obviously enhanced by forecasting volumes and locations jointly: for 80% of the users, the improvement is at most 10% of the prediction accuracy. The practical predictability of these two predictors is consistent with the theoretical predictability estimated by the joint entropy rate.

The last thing we evaluate in this section, is the practical predictability of data traffic volumes by knowing the previous locations of the same user, which we mark as $\pi(\mathcal{V})_{\texttt{PastL}}$. We portray in Figure 3(d) the CDF of the enhancement (*i.e.*, $\pi(\mathcal{V})_{\texttt{PastL}} - \pi(\mathcal{V})$) of each predictor by adding this mobility information as an additional knowledge in the prediction. As in the joint forecasting scenario shown in Figure 3(c), the Markovian predictors cannot benefit from this additional information (only 10% of the users have improvements), while the MLP predictor can have upto 10% of the improvement for 60% of the users. Although the rest of the users have reduced performance, we guess that their performance may be also improved by adding more data as historical information in prediction. Still, we can say that the predictors based on machine learning can utilize the mobility information efficiently in the prediction of mobile data traffic.

Consequently, our results regarding the joint practical predictability measured above is consistent with those of the joint theoretical predictability. We see that forecasting the data traffic volumes and locations jointly performs better than doing so separately, due to the spatiotemporal correlation of mobile data traffic. The problem is, to have such benefit in real-world prediction, we need machine learning techniques to better utilize the spatiotemporal correlation, while legacy Markovian methods are insufficient.

8 Conclusion

In this report, we analyze the predictability of per-user mobile data traffic, in isolation and jointly with mobility. Our data-driven analysis exploits the theoretical and practical performance in the prediction. In short, we conclude that there is a high degree of the predictability of individual mobile data traffic and it can be further enhanced by the knowledge of users' mobility. The main reason of the high predictability is the existence of the spatiotemporal correlation in each user's mobile data traffic dynamics. In real-world prediction, the legacy Markovian approach could achieve a fairly good prediction accuracy, while the key to further enhance the performance is the use of context information (*e.g.*, time of events or locations). For that, the novel machine learning techniques are quite useful. Nevertheless, there is still a gap between the real-world prediction accuracy and the theoretical accuracy upper bound.

References

[1] 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.

[2] W. Su, S.-J. Lee, and M. Gerla, "Mobility prediction in wireless networks," in *IEEE MILCOM 2000*, vol. 1, pp. 491–495, IEEE, 2000.

[3] P. N. Pathirana, A. V. Savkin, and S. Jha, "Mobility modelling and trajectory prediction for cellular networks with mobile base stations," in *ACM MobiHoc 2003c*, pp. 213–221, ACM, 2003.

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

[5] D. G. Taylor and M. Levin, "Predicting mobile app usage for purchasing and information-sharing," *International Journal of Retail & Distribution Management*, vol. 42, no. 8, pp. 759–774, 2014.

[6] W.-S. Soh and H. S. Kim, "Qos provisioning in cellular networks based on mobility prediction techniques," *IEEE Communications Magazine*, vol. 41, no. 1, pp. 86–92, 2003.

[7] H. Petander, "Energy-aware network selection using traffic estimation," in *ACM MICNET 2009*, pp. 55–60, ACM, 2009.

[8] V. A. Siris and D. Kalyvas, "Enhancing mobile data offloading with mobility prediction and prefetching," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 17, no. 1, pp. 22–29, 2013.

[9] Z. Li, J. Bi, and S. Chen, "Traffic prediction-based fast rerouting algorithm for wireless multimedia sensor networks," *International Journal of Distributed Sensor Networks*, 2013.

[10] G. Chen, S. Hoteit, A. Carneiro Viana, M. Fiore, and C. Sarraute, "Spatio-Temporal Predictability of Cellular Data Traffic," Research Report RT-0483, INRIA Saclay - Ile-de-France, Jan. 2017.

[11] G. L. Ulmer, *Internet invention: From literacy to electracy*. Longman New York, 2003.

[12] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks.," *INFOCOM*, pp. 882–890, 2011.

[13] S. Hoteit, S. Secci, Z. He, C. Ziemlicki, Z. Smoreda, C. Ratti, and G. Pujolle, "Content consumption cartography of the Paris urban region using cellular probe data," in *Proceedings of the First Workshop on Urban Networking*, UrbaNe '12, (New York, NY, USA), pp. 43–48, ACM, 2012.

[14] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, and Y. Li, "Big Data Driven Mobile Traffic Understanding and Forecasting: A Time Series Approach," *IEEE Transactions on Services Computing*, vol. 9, pp. 796–805, Aug. 2016.

[15] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," *IEEE/ACM Transactions on Networking (TON)*, vol. 25, pp. 1147–1161, Apr. 2017.

[16] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2016.

[17] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *ACM SIGMETRICS 2011*, pp. 305–316, ACM, 2011.

[18] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area.," *Pervasive and Mobile Computing*, pp. 230–235, 2015.

[19] A. Fumo, M. Fiore, and R. Stanica, "Joint spatial and temporal classification of mobile traffic demands," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, pp. 1–9, IEEE, 2017.

[20] I. Trestian, S. Ranjan, and A. Kuzmanovic, "Measuring serendipity: connecting people, locations and interests in a mobile 3G network," in *Proceedings of the 9th ...*, 2009.

[21] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *INFOCOM, 2012 Proceedings IEEE*, pp. 1341–1349, IEEE, 2012.

[22] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang, "The predictability of cellular networks traffic," in *2012 International Symposium on Communications and Information Technologies (ISCIT)*, pp. 973–978, IEEE, 2012.

[23] 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.

[24] Y. Zang, F. Ni, Z. Feng, S. Cui, and Z. Ding, "Wavelet transform processing for cellular traffic prediction in machine learning networks.," *ChinaSIP*, pp. 458–462, 2015.

[25] Z. Yi, X. Dong, X. Zhang, and W. W. 0007, "Spatial traffic prediction for wireless cellular system based on base stations social network.," *SysCon*, 2016.

[26] A. Nika, A. Ismail, B. Y. Zhao, S. Gaito, G. P. R. 0001, and H. Zheng, "Understanding and Predicting Data Hotspots in Cellular Networks.," *MONET*, vol. 21, no. 3, pp. 402–413, 2016.

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

[28] Y. Zhang and A. Årvidsson, "Understanding the characteristics of cellular data traffic," in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, pp. 13–18, ACM, 2012.

[29] R. Li, Z. Zhao, J. Zheng, Y. Chen, C. Mei, Y. Cai, and H. Zhang, "The Learning and Prediction of Application-level Traffic Data in Cellular Networks," *arXiv.org*, June 2016.

[30] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, pp. 180–186, ACM, 2017.

[31] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting whatsapp through large-scale measurements in mobile networks," in *ACM SIGCOMM Computer Communication Review*, vol. 44, pp. 133–134, ACM, 2014.

[32] Q. Deng, Z. Li, Q. Wu, C. Xu, and G. Xie, "An empirical study of the wechat mobile instant messaging service," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 390–395, May 2017.

[33] J. Erman, A. Gerber, K. Ramadrishnan, S. Sen, and O. Spatscheck, "Over the top video: the gorilla in cellular networks," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 127–136, ACM, 2011.

[34] Z. Li, X. Wang, N. Huang, M. A. Kaafar, Z. Li, J. Zhou, G. Xie, and P. Steenkiste, "An empirical analysis of a large-scale mobile cloud storage service," in *Proceedings of the 2016 Internet Measurement Conference*, pp. 287–301, ACM, 2016.

[35] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," *Computer Networks*, vol. 112, pp. 176–193, 2017.

[36] 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.

[37] Y. Li, J. Yang, and N. Ansari, "Cellular smartphone traffic and user behavior analysis," in *ICC 2014 - 2014 IEEE International Conference on Communications*, pp. 1326–1331, IEEE, 2014.

[38] N. Bui, F. Michelinakis, and J. Widmer, "A model for throughput prediction for mobile users," *European Wireless 2014; 20th ...*, 2014.

[39] N. Bui and J. Widmer, "Modelling Throughput Prediction Errors as Gaussian Random Walks," Sept. 2014.

[40] B. Liu, D. Niu, Z. Li, and H. V. Zhao, "Network latency prediction for personal devices: Distance-feature decomposition from 3D sampling," in *IEEE INFOCOM 2015 - IEEE Conference on Computer Communications*, pp. 307–315, IEEE, 2015.

[41] R. Zhu, B. Liu, D. Niu, Z. Li, and H. V. Zhao, "Network Latency Estimation for Personal Devices - A Matrix Completion Approach.," *IEEE/ACM Trans. Netw.*, 2017.

[42] G. Chen, S. Hoteit, A. Carneiro Viana, M. Fiore, and C. Sarraute, "Enriching Sparse Mobility Information in Call Detail Records," Technical Report RT-0496, INRIA Saclay - Ile-de-France, Nov. 2017.

[43] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[44] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive Wi-Fi mobility data," *IEEE Transactions on Mobile ...*, 2006.

[45] J. Jeong, M. Leconte, and A. Proutiere, "Cluster-aided mobility predictions," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, IEEE, Apr. 2016.

[46] A. Moffat, "Implementing the ppm data compression scheme," *IEEE Transactions on communications*, vol. 38, no. 11, pp. 1917–1921, 1990.

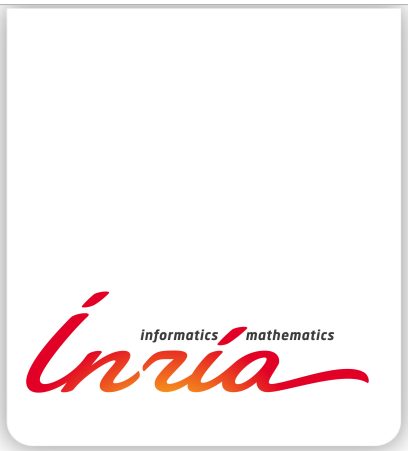
[47] P. Jacquet, W. Szpankowski, and I. Apostol, "An universal predictor based on pattern matching, preliminary results," *Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities*, pp. 75–85, 2000.

[48] K. Gopalratnam and D. J. Cook, "Active lezi: An incremental parsing algorithm for sequential prediction," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 04, pp. 917–929, 2004.

[49] "Ensemble methods – scikit-learn." `http://scikit-learn.org/stable/modules/ensemble.html`.

[50] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803