Spatio-Temporal Predictability of Cellular Data Traffic

Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore, Carlos Sarraute

Spatio-Temporal Predictability of Cellular Data Traffic

Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore, Carlos Sarraute

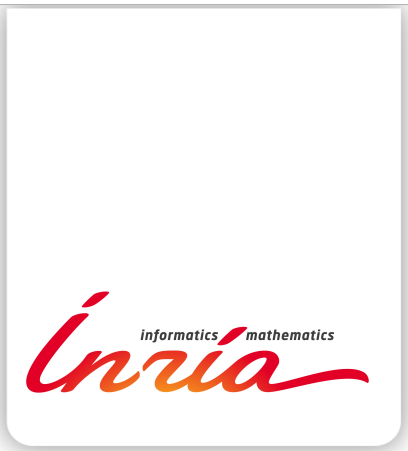
TECHNICAL REPORT

N° 483

November 2016

Project-Teams INFINE

ISSN 0249-0803 ISRN INRIA/RT--483--FR+ENG



Spatio-Temporal Predictability of Cellular Data Traffic

Guangshuo Chen*†, Sahar Hoteit‡, Aline Carneiro Viana†,

Marco Fiore§, Carlos Sarraute¶

Project-Teams INFINE

Technical Report n° 483 — November 2016 — 17 pages

Abstract: The ability to foresee the data traffic activity of subscribers opens new opportunities to reshape mobile network management and services. In this paper, we leverage two large-scale real-world datasets collected by a major mobile carrier in Mexico to study how predictable are the cellular data traffic demands generated by individual users. We first focus on the predictability of mobile traffic consumption patterns in isolation. Our results show that it is possible to anticipate the individual demand with a typical accuracy of 85%, and reveal that this percentage is consistent across all user types. Despite the heterogeneity in usage patterns of users, we also find a lack of significant variability in predictability when considering demographic factors or different mobility or mobile service usage. Then, we analyze the joint predictability of the traffic demands and mobility patterns. We find that the two dimensions are correlated, which improves the predictability upper bound to 90% on average.

Key-words: user profiling and personalization

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

* Université Paris Saclay, France
† INRIA Saclay, France
‡ Ecole d'ingénieurs du numérique ISEP, France
§ CNR - IEIIT, Italy
¶ Grandata Labs, Argentina

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

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

Prévisibilité spatio-temporelle du trafic de données cellulaires

Résumé : La capacité de prévoir l'activité des abonnés par rapport à leur trafic de données ouvre des nouvelles possibilités de remodeler la gestion et les services de réseaux mobiles. Dans cet article, nous exploitons deux grands ensembles de données réels collectés par un important opérateur mobile au Mexique afin d'étudier la prévisibilité des demandes de trafic de données cellulaires générées par les utilisateurs, de manière individuelle. Nous nous concentrons d'abord sur la prévisibilité des modèles de consommation de trafic mobile isolément. Nos résultats montrent qu'il est possible d'anticiper la demande individuelle avec une précision typique de 85%, et de révéler que ce pourcentage est cohérent pour tous les types d'utilisateurs. Malgré l'hétérogénéité des demandes des utilisateurs, nous constatons également une absence de variabilité significative de la prévisibilité en tenant compte des facteurs démographiques, de la mobilité, ou de l'utilisation des services mobiles. Ensuite, nous analysons la prévisibilité des demandes de trafic combinée à des modèles de mobilité. Nous constatons que les deux dimensions sont corrélées, ce qui améliore la limite supérieure de prévisibilité à 90% en moyenne.

Mots-clés : profilage et personnalisation de l'utilisateur

1 Introduction

The quantitative understanding of human behavior has recently emerged as a central question in multi-disciplinary research [1, 2, 3, 4, 5]. Individual actions are the root cause of dynamics that impact technological and economic phenomena of interest to many research communities. In wireless networking, the ability to foresee human activities has important implications, *e.g.*, in resource management or service provisioning [6]. The performance of any practical technique that aims at anticipating such human behaviors is bounded by its theoretical *predictability*, which evaluates to what degree a specific behavior can be foreseen. For instance, human mobility was found to be highly predictable [3].

Previous studies have shown that there exists some regularity in the mobile data traffic activity of individual subscribers [7, 8, 9, 10]. However, there is no analysis of how such per-user regularity is translated into actual predictability: this kind of study was only carried out for the aggregate traffic load recorded at cellular base stations [5, 11].

In this paper, we aim at filling the gap above, and provide a first investigation of the predictability of the volume of mobile data traffic generated by individual users. Our study allows answering an important question: *to what degree is the individual consumption of mobile data traffic predictable?* To that end, we mine two large-scale real-world datasets describing the cellular communication activity of thousands of subscribers, and leverage tools from information theory to determine predictability bounds. Our contributions are summarized as follows.

- We provide a first study of the predictability of mobile data traffic usage from the viewpoint of individual subscribers. We find that, by just considering temporal correlations in the traffic, 85% 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.
- We extend the methodology so as to account for the joint predictability of single users' traffic consumption and movement patterns. This let us investigate whether it is possible to forecast when, where, and how much mobile data traffic is generated by individual subscribers.
- We observe a 90% potential predictability of the spatio-temporal data consumption patterns of individual users. This result is due to the strong correlation between mobility and mobile service usage, *i.e.*, to the fact that subscribers tend to generate similar amounts of traffic at each location.

The rest of the paper is organized as follows. Sec. 2 sheds light on the related works from the literature. In Sec. 3, we introduce the concept of predictability and the methodology to compute it. In Sec. 4, we present our datasets and discuss data preprocessing. Sec. 5 discusses the predictability of mobile data traffic in isolation. Sec. 6 extends the analysis to the joint predictability of mobility and data traffic. Finally, Sec. 7 concludes the paper.

2 Related work

Since early 2000s, traffic predictability has attracted much attention in the wired networking community [12, 13, 14]. Nevertheless, after a decade, network environment and most importantly, the type of traffic users are generating have significantly changed: Users have become eager to engage with mobile applications and connected services. These brought the necessity to reconsider the data traffic analysis in the context of wireless or cellular networks.

As a consequence, the literature on the study of cellular network traffic has grown dramatically [15]. In this context, the characterization of mobile data traffic, *i.e.*, subscribers' data consumption, is of fundamental importance, as it impacts the design of solutions for network load balancing and aims at improving the quality of Internet-based mobile services. In this work, we focus on the prediction of individual subscribers' traffic data volume consumption, which allows network operators to manage their resources in advance, so as to accommodate the future demand at lower maintenance and operational costs. It is also important in defining traffic plans that are better tailored to users' needs.

A number of studies have attempted to understand cellular data traffic. The authors in [16] modeled the volume distribution of Internet data traffic towards an improved traffic volume prediction. Oliveira *et al.* [17] proposed a measurement-driven model of mobile data traffic and a synthetic mobile data traffic generator. Both these studies do not consider the influence of subscribers' mobility on their mobile service consumption. The relation between content consumption and mobility properties is considered in studies that focus on application interests [18], data traffic dynamics [19] and service usages [8]. However, none of these works provides a complete analysis of the predictability of the mobile data traffic consumption of individual subscribers, with respect to both time and space dimensions – as done in this work.

Recently, two studies contributed to our understanding of the predictability of cellular data traffic. Zhou *et al.* [5] analyzed the predictability of voice, text and data traffic in cellular networks. Li *et al.* [11] focused on the traffic predictability in Cloud radio access networks (C-RANs), and proposed future potential of software-defined C-RAN paradigms that benefit from the traffic prediction. These two works have an aggregated perspective, only distinguishing all traffic served by each base station. Instead, we analyze the traffic predictability from a more challenging viewpoint of individual users.

Our analysis is driven by information theory [20]. Tools of entropy estimation are introduced in [21, 22], including the entropy estimator we used in the paper. Entropy is a concept found in many studies investigating spatial or temporal characteristics of cellular network data [23, 16, 4, 18]. Our work is mainly based on the research of Song *et al.* who were the first to use information theory to investigate the predictability of human mobility [3, 24]. However, the goal of our study is different, as the target is mobile data traffic volume predictability, studied in isolation as well as jointly with mobility.

3 Measuring predictability

In this section, we define the concept of *predictability* as employed in our work, and show how its *upper limit* can be derived from *entropy*. Our methodology follows that first proposed in [3], which has been successfully employed to investigate the predictability of human mobility [3, 25], mobile data traffic [5], and vehicular traffic [26]. To have a quantitative manner, we favor the definition of predictability in [3] over equivalent ones (*e.g.*, those in [27] and [12]), since it is more easily adapted to the study of the joint predictability of mobility and mobile data traffic demands, which is our ultimate objective.

3.1 Predictability

Consider a particular human behavior (*e.g.*, a user's whereabouts or data traffic volumes), which could be symbolized discretely and be measured regularly at every time interval. Its *predictability* at the i^{th} time interval, denoted as Π_i, is defined as the maximum probability of correctly forecasting the current state from a known set of possible outcomes. Leveraging the concept of expected value, the overall predictability is then defined as:

$$\Pi \equiv \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \Pi_i. \tag{1}$$

As indicated in Sec. 1, we are interested in investigating the predictability of the volume of data traffic generated by mobile phone subscribers, in isolation as well as jointly with the geographical locations where they consume the associated mobile services. As shown in Eqn. 1, the computation of the overall predictability requires knowledge of the expected predictability over an infinite time interval. Since measurement periods are finite in any practical setting, we model user-generated data traffic demands and user movements as stochastic processes. We then compute an upper bound on each of their predictability. To that end, we employ the empirical estimation proposed in [3] and its supplement [24], presented next.

3.2 An entropy-based upper bound of predictability

In information theory [20], entropy measures the degree of uncertainty or disorder of an information flow. Intuitively, entropy and predictability are negatively correlated variables: a random process with low (or high) uncertainty is highly (or little) predictable. Song *et al.* [3, 24] studied this correlation and established an explicit formula from the intuition. Their formula quantifies the correlation between entropy $H(X)$ (or entropy rate $\mathcal{H}(\mathbb{X})$) and an upper bound on predictability $\Pi^{\max}$, and is summarized as follows.

Recall the human behavior mentioned above. Let discrete random variable X_i denotes the symbolized behavior at the i^{th} time interval with probability mass function $p_i(x) = \Pr(X_i = x)$. Its *entropy* is formulated as $H(X_i) \equiv -\sum_{x_i \in X_i} p_i(x_i) \log_2 p_i(x_i)$ [20]. Its predictability, denoted as Π_i, is linked with its entropy by the inequality as $H(X_i) \leq H_F(\Pi_i, N_i)$ where N_i is the number of possible states at the i^{th} time interval and $H_F(p, N)$ is a function defined as:

$$H_F(p, N) = p \log_2 p + (1 - p) \log_2 \frac{1 - p}{N - 1}. \tag{2}$$

Because $H_F(p, N)$ is monotonically increasing, the upper bound of Π_i is derived from the inequality as $\Pi_i \leq \Pi_i^{\max}$ where $\Pi_i^{\max}$ satisfies $H(X_i) = H_F(\Pi_i^{\max}, N_i)$: We could calculate an upper bound of the behavior's predictability if we know its entropy at the i^{th} interval. According to [24], this upper bound estimation is *tight*: It could be possibly achieved by an actual algorithm.

Consider a time series of behaviors, denoted as $\mathbb{X} = \{X_1, \cdots, X_T\}$. Similarly, its *entropy rate* is defined, regarding to the joint entropy of all behaviors over T time intervals in $\mathbb{X}$, as follows:

$$\mathcal{H}(\mathbb{X}) \equiv \lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} H(X_i | X_{i-1}, \cdots, X_1). \tag{3}$$

It could be regarded as *per-symbol* entropy describing the mean degree of uncertainty of each behavior given the condition of being aware of past ones. As a bridge between its entropy rate and its overall predictability, the following inequality stands: $\mathcal{H}(\mathbb{X}) \leq H_F(\Pi, N)$ where N is the number of all possible states which may appear in the time series. From this inequality, an upper limit of the overall predictability is also derived as $\Pi \leq \Pi^{\max}$. Apparently, the upper bound $\Pi^{\max}$ is the exclusive solution of the equation $H_F(\Pi^{\max}, N) = \mathcal{H}(\mathbb{X})$.

In our work, human behaviors are analyzed on per-user basis. Thus, given a particular behavior of a user u, we collect a time series of the behavior as $\mathbb{S}_u$, and calculate its entropy rate and upper limit of the overall predictability. Because $\mathbb{S}_u$ is a limited set and consequently, is unsuitable to apply Eqn. 3, the entropy rate is empirically estimated on the Lempel-Ziv encoding [22] as:

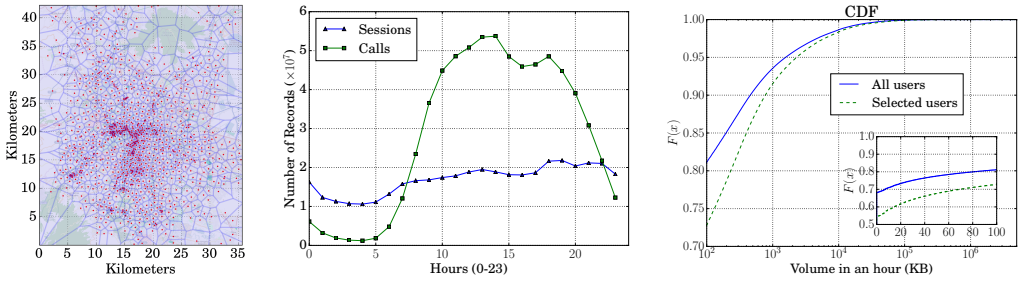
$$\mathcal{H}(\mathbb{X}) \hat{=} \mathcal{H}^{est}(\mathbb{S}_u) \equiv \frac{T \log_2 T}{\sum_{i=1}^{T} L^i}, \tag{4}$$

where T is the total number of time intervals and L_i is the length of the shortest subsequence beginning from i which never appears before the i^{th} interval. Theoretically, $\mathcal{H}^{est}(\mathbb{S}_u)$ converges to $\mathcal{H}(\mathbb{X})$ with $T \to \infty$ [22]. For the overall predictability, we calculate its upper bound by applying a numerical solver on the corresponding equation. The solver receives N and $\mathcal{H}(\mathbb{X})$ as inputs and produces the upper limit.

4 Data overview

Our study is based on two real-world anonymized datasets describing the cellular network activity of $1,598,329$ mobile phone subscribers (identically called users) of a major 3G cellular operator in Mexico City. All data refer to a 3-month period, from October 1 to December 31, 2014. 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



(a) Cell locations (b) Number of records (c) Volume of mobile data

Figure 1: (a) Deployment of cell towers in Mexico City. (b) Number of CDR (green curve) and Internet data session (blue curve) records per hour. (c) CDF of the hourly data traffic volume generated by all mobile phone subscribers (solid blue curve), and by the selected 45,832 users (dashed green curve).

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.

4.1 Datasets

4.1.1 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.

From a spatial perspective, cell tower locations are fairly dense in Mexico City, as shown in Fig. 1(a), where (red) dots represent the base stations and the Voronoi tessellation approximates the coverage of each cell: on average, a cell tower covers a 2 km^2 area. From a temporal perspective, the frequency of CDR entries (*i.e.*, of phone calls) is not uniform over time. Fig. 1(b) shows that users in Mexico City are more keen to make or receive calls during daytime, with an activity peak around midday.

4.1.2 Internet data 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 [28]. Each entry in the dataset contains the hashed device identifier, the volume of upload and download data exchanged in KiloBytes, and the timestamp denoting the starting time of the session. The device identifiers are shared by both datasets and are hashed cryptographically.

The dataset does not contain spatial information, but, from a temporal perspective, data sessions have a relatively uniform pattern, shown in Fig. 1(b). This is quite different from voice calls and is mostly due to automatically-generated background traffic and push notifications that are periodic and fairly independent of human activity.

To give a clear view of the mobile data traffic demands in the dataset, Fig. 1(c) portrays the cumulative distribution function (CDF) of the traffic volume generated by each subscriber on an hourly basis (solid blue curve). We remark that: *(i)* in 68% of hours, a subscriber does not start any new data session; *(ii)* the hourly traffic is highly heterogeneous, and varies from 1 KB to 4.6 GB; *(iii)* only 6% of hours present a traffic volume higher than 1 MB/h.

4.2 Data preprocessing

As detailed in Sec. 3, the calculation of the entropy rate and the corresponding upper bound on predictability requires symbolized versions of the movements and demands of mobile phone subscribers. To this end, we proceed as follows. Firstly, we complete the CDR dataset to increase the number of locations identified during nighttime. Secondly, we filter subscribers in the datasets

so as to ensure that their mobile activities allow accurate entropy estimation. Finally, we merge the CDR and session logs for the selected subscribers, and compute representative symbolized time series of locations $\mathbb{S}_u^{loc}$ and of data traffic volumes $\mathbb{S}_u^{vol}$.

4.2.1 Completion of CDR dataset

:

It is well-known that the events logged in CDR data tend to be sparse in time. Moreover, they are not uniformly distributed, as seen in Fig. 1(b). In order to improve the accuracy of the user location information in the CDR data, we employ the recent `stop-by-spothome` completion technique [29]. The `stop-by-spothome` approach allows increasing the temporal coverage of CDR without affecting the localization precision. Formally, for each subscriber $u \in \mathcal{U}$, where $\mathcal{U}$ is the observed population, `stop-by-spothome` performs the following operations.

- The position of user u is considered to be the same associated to a CDR data entry (*i.e.*, when a voice or sms event is triggered) 30 minutes before to 30 minutes after the entry timestamp.
- If two CDR entries are at less than 1 hour distance, then the transition from the location associated to the first entry and that associated to the second entry occurs (instantaneously) halfway between the two CDR entry timestamps.
- The home location ℓ_u^H is determined as the cell where u is most frequently found during the night time interval $t^H = (22h, 9h)$, according to the CDR data.
- The home boundary time $t_u^H \subseteq t^H$ is then defined as the most probable interval during which the subscriber is found at ℓ_u^H in the CDR data.
- If a subscriber u's location at a time instant $t \in t_u^H$ is unknown and if he was last seen at no more than 1 km from this home location ℓ_u^H, he is considered to be at ℓ_u^H at time t.

The technique has been validated against ground-truth GPS data, and we refer the reader to [29] for all details, including extensive validation and comparative evaluation.

4.2.2 User filtering

To ensure statistical significance of the analysis, subscribers need to be sufficiently active during the period under study. An exceedingly small number of voice call or data session activities risks to lead to considerably incomplete mobility information or negligible mobile data traffic demand. Therefore, we focus on users whose movements can be tracked efficiently via CDRs, and who consistently use the cellular network for Internet access.

We calculate the daily volume of mobile data traffic generated by each subscriber in the *Internet session dataset*, and filter out erratic subscribers who establish sessions in less than 73 days (*i.e.*, 80% of the measurement period). Secondly, we compute, for each subscriber, the *incompleteness* q, *i.e.*, the fraction of hours during which no positioning information is available from her CDR entries. We then filter out subscribers with $q \geq 0.8$ or who only visited 2 locations at most during the measurement period. The secondary conditions have been proven instrumental to an accurate entropy estimation [3].

As a result of our filtering process, 45,832 subscribers are retained for our analysis. The green dashed curve in Fig. 1(c) shows the CDF of the mobile data traffic volume generated by the selected subscribers. Despite a minor bias in the fraction of hours during which no session is established (*i.e.*, this fraction decreases from 68% to 55%), the distribution is fairly consistent with that of the complete user set. In particular, the strong heterogeneity that characterizes the demand of all subscribers is still well present in the selected user subset.

4.2.3 Volume integration

We construct symbolized time series of the mobile data traffic generated by a subscriber $u \in \mathcal{U}$ as:

$$\mathbb{S}_u^{vol} = \{v_u^1, v_u^2, \ldots, v_u^i, \ldots, v_u^T\}, \tag{5}$$

where v_u^i is a measure of the traffic volume generated by the subscriber u during the i^{th} time interval. We consider four different temporal resolutions for time intervals: 15, 30, 45, or 60 minutes. In each case, we aggregate the volume of traffic from all sessions recorded during the corresponding interval. For instance, when the resolution is 60 minutes, the measurement period is divided into $T = 24$ hours/day $\times$ 92 days $= 2208$ intervals, and the mobile data traffic is computed on an hourly basis: This will be our default setting, unless stated otherwise.

As seen in Fig. 1(c), such traffic varies across a wide spectrum, from KB/h to GB/h. Since the traffic volume needs to be quantized in the $\mathbb{S}_u^{vol}$ representation, we favor a representation that captures the traffic magnitude over a uniform discretization. The rationale is that one is more interested in predicting whether a user will generate, e.g., 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 following five different quantizations of the traffic volume spectrum, listed in order of increasing accuracy:

- **Q1**: four quantization levels, *i.e.*, *idle* (0 KB), *light* (1 KB, 1 MB), *heavy* (1 MB, 1 GB), and *extremely heavy* (1 GB, 10 GB).
- **Q2**: eight quantization levels, *i.e.*, 0, $(1, 10)$, $(10, 10^2)$, ..., $(10^6, 10^7)$, all values in KB. Once stated otherwise, this is our default setting.
- **Q3**: twelve quantization levels, obtained by bisecting each level over 1 MB in Q2, *e.g.*, splitting (1,10) MB into (1,5.5) MB and (5.5,10) MB.
- **Q4**: sixteen quantization levels, obtained by trisecting each level over 1 MB in Q2.
- **Q5**: forty quantization levels, obtained by nine-secting each level over 1 MB in Q2.

4.2.4 Location integration

The movement of a user $u \in \mathcal{U}$ is represented as a symbolized time series of locations, as follows:

$$\mathbb{S}_u^{loc} = \{\ell_u^1, \ell_u^2, \ldots, \ell_u^i, \ldots, \ell_u^T\}, \tag{6}$$

where ℓ_u^i is the location of u in the i^{th} interval spanning (t_{start}^i, t_{end}^i). The location is that of the cell to which the user u is attached most during the period (t_{start}^i, t_{end}^i). If that location is unidentified, *i.e.*, no entry is available in the CDR dataset for u during (t_{start}^i, t_{end}^i), ℓ_u^i is marked as *unknown*. Since the CDR logs are sparse in time, we only implement the temporal resolution of 60 minutes for time intervals. For that, the same measurement period as that of mobile data traffic is split into the $T = 24$ hours/day $\times$ 92 days $= 2208$ hour-long intervals and each representative location is selected on an hourly basis.

5 Predictability of mobile data traffic

In this section, we study the predictability of the mobile data traffic generated by individual subscribers. For now, we focus on the forecast of traffic volume in isolation, and we will consider the joint predictability of traffic and mobility later on.

5.1 Methodology

We implement the method presented in Sec. 3 to empirically derive an upper bound on the predictability through the entropy rate. Formally, we consider a stochastic process $\mathbb{V} = \{V_1, \cdots, V_T\}$ that describes the mobile demand of a generic subscriber as a sequence of symbolized traffic volumes V_i, for each time interval i.

The *actual entropy* rate, denoted by $\mathcal{H}(\mathbb{V})$, depends not only on the frequency of appearance of each symbolized traffic volume but also on the order in which they appear, capturing the temporal order presented in a subscriber's traffic usage pattern. Formally, $\mathcal{H}(\mathbb{V})$ is defined in Eqn. 3 and models the process as a stationary stochastic process. For each user, an empirical estimation of $\mathcal{H}(\mathbb{V})$ is calculated from $\mathbb{S}_u^{vol}$ as described in Sec. 3. We find the estimator to

converge rapidly enough (within a few days in our three-month data) in the case of practical traffic volumes.

We leverage $\mathbb{S}_u^{vol}$ to derive three additional variants of the entropy rate, which we will use to investigate the properties of the process. The variants are as follows:

- The *random entropy* is computed by considering that each V_i is equally probable and time-independent in the process. The random entropy is $H^{\text{rand}}(V) \equiv \log_2 N$, where N is the number of distinct quantizations of traffic volumes associated to a subscriber. It indicates the theoretical maximum value of the entropy rate $\mathcal{H}(\mathbb{V})$.

- The *temporal-uncorrelated entropy* is formulated as $H^{\text{unc}}(V) \equiv -\sum_{v \in V} p(v) \log_2 p(v)$, where $p(v)$ is derived from $\mathbb{S}_u^{vol}$ and represents the historical probability of a subscriber to generate volume v, characterizing the heterogeneity of traffic demand patterns. The temporal-uncorrelated entropy characterizes a mobile demand process that has no temporal correlations, hence its name.

- The *nonzero-temporal-uncorrelated entropy* is based on the same model of $H^{\text{unc}}(V)$, but it is limited to those cases when the user is not idle. Formally, it is $H^{\text{n0}}(V) \equiv -\sum_{v \in V/\{0\}} p(v|v \neq 0) \log_2 p(v|v \neq 0)$. The nonzero-temporal-uncorrelated entropy captures the heterogeneity of the traffic volume exchanged during active sessions only, assuming again that no temporal correlations exist among them.

It is worth noting that, naturally, for each subscriber: $\mathcal{H}(\mathbb{V}) \leq H^{\text{unc}}(V) \leq H^{\text{rand}}(V)$.

As discussed in Sec. 3, an upper bound on the predictability can be computed from the entropy rate. In our context, this bound is an estimation of the maximum achievable accuracy in the prediction of the mobile traffic demand, given a particular model of the distribution of $\mathbb{V}$. Hence, four upper bounds for the predictability, indicated as $\Pi^{\text{rand}}(V)$, $\Pi^{\text{unc}}(V)$, $\Pi^{\text{n0}}(V)$ and $\Pi^{\max}(\mathbb{V})$, can be calculated from the entropy rates above.

In addition, recall that multiple representations of $\mathbb{S}_u^{vol}$ are possible, depending on the combination of time granularity and volume quantization. For each such combination, different results are obtained in terms of entropy rate and thus predictability.

5.2 Baseline results

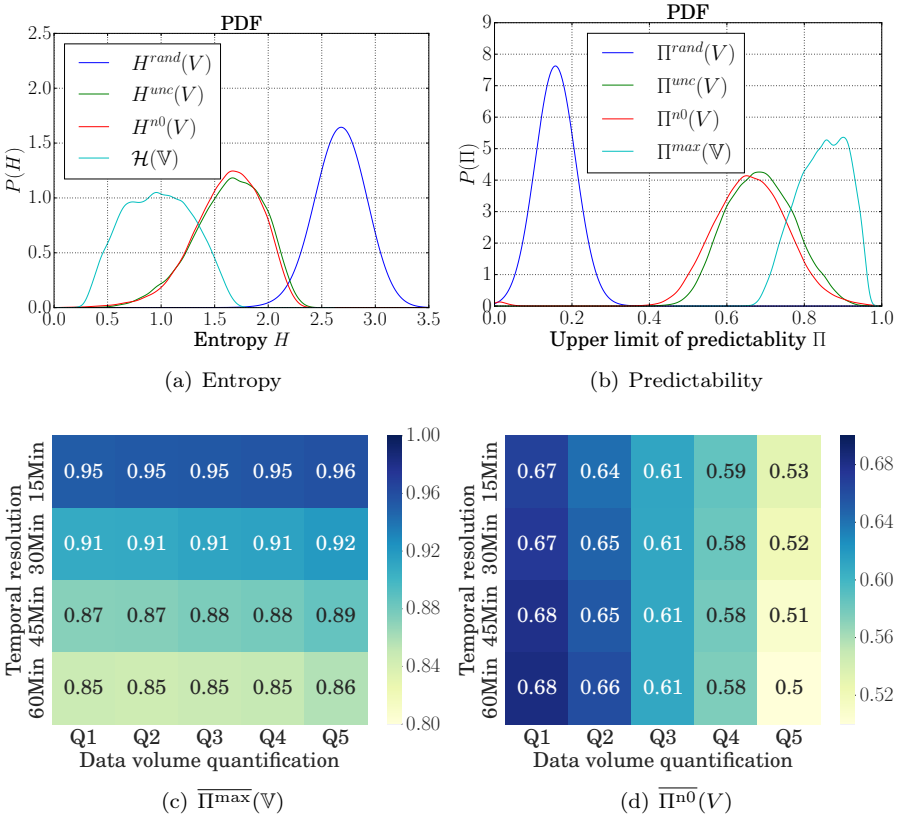
Our baseline results are shown in Fig. 2, and are obtained under 1-hour (time) and Q2 (traffic volume) quantization level described in Sec. 4.2.3. Specifically, Fig. 2(a) displays the probability density function (PDF)[1] of the four versions of the entropy rate presented in Sec. 5.1. Fig. 2(b) portrays the PDF of the corresponding upper bounds on predictability.

Let us start by considering the PDF of $H^{\text{rand}}(V)$ in Fig. 2(a). Its range indicates that an equiprobable distribution of traffic volume during each time interval can be represented with three bits. This phenomenon is normal, as we consider eight traffic volume quantization levels as our default setting. When the temporal-uncorrelated entropy, i.e., $H^{\text{unc}}(V)$, is concerned, a sizable shift of probability occurs. The uncertainty decreases to $2^{H^{\text{unc}}(V)} = 2^{1.63} \approx 3$. Under this model, each user tends to generate traffic that is described by just three quantization levels out of the eight available. For instance, at each time interval, a user may generate traffic by one order of MB or tens of MB, or stay idle; but typically she will not generate smaller or larger traffic volumes. The same holds for users who generate, e.g., order-of-KB or order-of-GB traffic. Ultimately, a reduced entropy rate implies higher regularity in the mobile traffic demand.

Interestingly, idle time intervals do not bias such regularity. Indeed, the PDF of $H^{\text{n0}}(V)$ overlaps well to that of $H^{\text{unc}}(V)$, suggesting that the considerations above also hold when only time intervals with data sessions are considered. However, our main result is the significant shift presented by the PDF of $\mathcal{H}(\mathbb{V})$, which is amassed around a value 0.97. When taking the temporal ordering of data sessions into account, one can reduce the uncertainty to just two quantization levels.

The probability distributions in Fig. 2(b) confirm these findings and provide upper numerical bounds to the predictability of the mobile data traffic demand generated by individual subscribers. We observe that $\Pi^{\text{rand}}(V)$ peaks at 0.16, i.e., it is very hard to guess the volume of

[1] All PDFs are represented using *kernel density estimation* (KDE).



(a) Entropy (b) Predictability

(c) $\overline{\Pi^{\max}}(\mathbb{V})$ (d) $\overline{\Pi^{n0}}(V)$

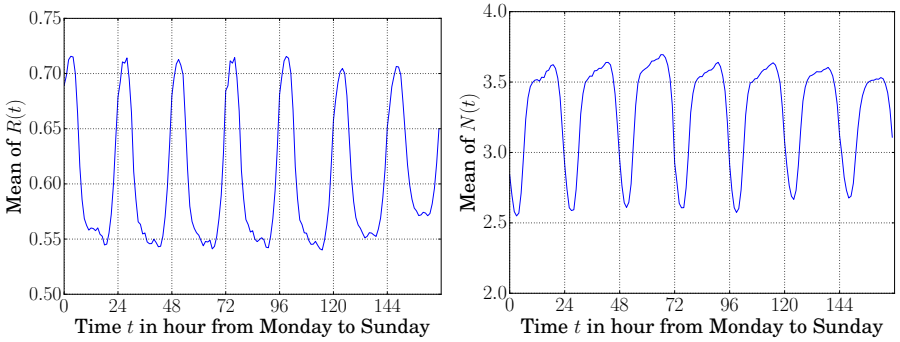
Figure 2: (a) Distributions of the random entropy $H^{\text{rand}}(V)$, the temporal-uncorrelated entropy $H^{\text{unc}}(V)$, the nonzero-temporal-uncorrelated entropy $H^{\text{n0}}(V)$, and the entropy rate $\mathcal{H}(\mathbb{V})$ as observed in the individual traffic demand generated by the selected 45,832 users. (b) Equivalent distributions of the upper bounds on the predictability $\Pi^{\text{rand}}(V)$, $\Pi^{\text{unc}}(V)$, $\Pi^{\text{n0}}(V)$ and $\Pi^{\max}(\mathbb{V})$. (c) Heatmap of the median predictability upper bound $\Pi^{\max}(\mathbb{V})$ for different quantizations of time and traffic volume. (d) Heatmap of the median predictability upper bound $\Pi^{\text{n0}}(V)$ for different quantizations of time and traffic volume.

traffic generated by such a stochastic model. The predictability grows for $\Pi^{\text{unc}}(V)$ and $\Pi^{\text{n0}}(V)$, which peak at 0.69 and 0.66, respectively. More importantly, $\Pi^{\max}(\mathbb{V})$ indicates that the demand of a subscriber can be possibly predicted within 85% accuracy on average. It means that in only 15% of the time does the subscriber generate a traffic volume in a manner that appears to be random, but in the remaining 85% of the time, we could hope to predict her volume. This result proves, for the first time, that *not only the mobility of subscribers is highly predictable, but also the traffic volume that they generate via their mobile devices can be accurately anticipated.*

The results in Fig. 2(a) and Fig. 2(b) refer to the case where the individual mobile data traffic is represented with a temporal resolution of 60 minutes and volume quantization Q2. In fact, data granularity has been shown to have a significant impact in the predictability of mobility [30]. We thus explore if the same is true in the case of predictability of the mobile data traffic demand.

The heatmaps in Fig. 2(c) and in Fig. 2(d) show the median of the upper bound on the mobile demand predictability, over four temporal and five traffic volume quantization levels. The two plots refer to $\Pi^{\max}(\mathbb{V})$ and $\Pi^{\text{n0}}(V)$, respectively.

In Fig. 2(c), we observe that $\Pi^{\max}(\mathbb{V})$ is not significantly affected by the traffic volume quantization, i.e., our results appear to have general validity under different levels of accuracy in the representation of the mobile demand. In contrast, surprisingly, it grows with finer-grained temporal resolutions. The reason is that more idle intervals appear as the temporal resolution is increased; these idle intervals tend to dominate the real-world distribution of mobile data traffic, reducing the entropy and improving the overall predictability but hiding the predictability of non-idle intervals. This is confirmed by Fig. 2(d): $\Pi^{\text{n0}}(V)$, which only accounts for non-idle time intervals, is slightly affected by variations in the time granularity. Instead, it is strongly



(a) (b)

Figure 3: Temporal dynamics of individual mobile data traffic volume, during the average week. (a) Regularly $R(t)$. (b) Number of observed levels $N(t)$.

dependent on the traffic volume quantization, an artifact of the lack of temporal correlation in this model, which disappears in $\Pi^{\max}(\mathbb{V})$.

5.3 Temporal variability

It is well-known that the demand generated by mobile phone subscribers is time-dependent [17]. Thus, a relevant question is whether the predictability the mobile data traffic similarly undergoes temporal variations.

To that end, we compute, for each user and on an hourly basis, the regularity $R(t)$, *i.e.*, the probability that the user generates the most likely traffic volume observed during each hour. Similarly, we define $N(t)$ as the number of unique traffic volume levels observed at each hour. Regularity provides a lower bound on the predictability, as it ignores the temporal correlation of subscribers' traffic demand patterns [3, 24], and is typically inversely correlated with $N(t)$.

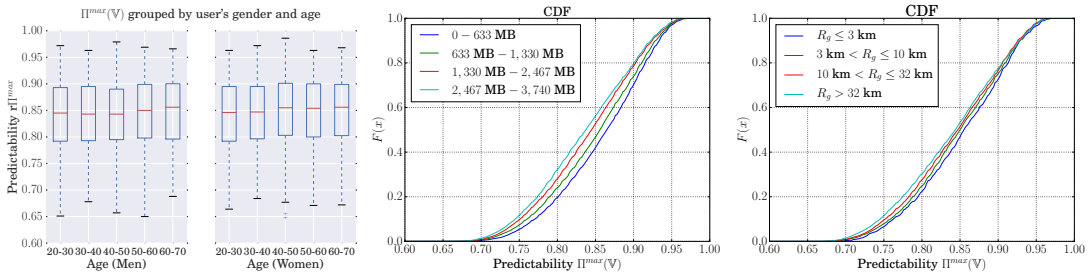
Fig. 3 shows the evolution of $R(t)$ and $N(t)$ over the average week. We remark a clear circadian rhythm. At night, the mean of $R(t)$ rises to approximately 0.7, meaning that, on average, the subscriber's demand matches the most likely traffic volume around 70% of the time. During the morning, working hours, and evening, $R(t)$ drops to 0.55. $N(t)$ shows opposite trends, as expected. As $R(t)$ and $N(t)$, as tied to the predictability, we conclude that the latter varies significantly over time as well. However, we do not observe significant variations from one day to another, which suggests that mobile data traffic volume predictability is not only imposed by the working schedule but is intrinsic to more generic human activities.

5.4 Variability across subscriber types

Our datasets allow us to explore several additional dimensions of the traffic volume predictability. These dimensions are related to the nature of the subscriber. The results are shown in Fig. 4 and discussed below.

- *Age and gender.* The age and gender of the mobile user are known to affect the way mobile services are consumed. However, Fig. 4(a) shows that these do not affect in a remarkable manner the predictability of the traffic volume. Hence, age- and gender-induced behaviors remain similarly predictable when it comes to the traffic volume.

- *Overall mobile data traffic volume consumption.* We categorize users into four groups, according to their data consumption during the 92 days of the data collection period. Each group consists of 25% of the observed users. As shown in Fig. 4(b), the predictability $\Pi^{\max}(\mathbb{V})$ tend to decrease as the data volume increases. Yet, the mean of $\Pi^{\max}(\mathbb{V})$ is 85.5% in the group of $0-674$ MB and 83.1% in the group of $2,553$ MB $-3,740$ MB. The difference is thus small and can be imputed to the fact that a larger amount of data naturally entails more complex dynamics. We conclude that the overall amount of generated traffic only marginally impacts the potential predictability of traffic volumes.



(a) (b) (c)

Figure 4: (a) Boxplot of $\Pi^{\max}(\mathbb{V})$ categorized by user's age and gender. Each box denotes the median, $25^{th} - 75^{th}$ percentiles, and minimum and maximum values. (b) CDF of $\Pi^{\max}(\mathbb{V})$, when users are separated according to the overall mobile data traffic volume recorded they generate during the measurement period of 92 days. Four groups are considered: $0 - 633$ MB, 633 MB $- 1,330$ MB, $1,330$ MB $- 2,467$ MB, and $2,467$ MB $- 3,740$ MB. Each group contains 25% of the observed users. (c) CDF of $\Pi^{\max}(\mathbb{V})$, when users are separated according to their level of mobility. Four ranges of radius of gyration are considered, mapping to sedentary ($R_g \leq 3$km), urban (3 km $< R_g \leq 10$ km), peri-urban (10 km $< R_g \leq 32$ km), and long-range commuting ($R_g > 32$ km) profiles.

- *Mobility level.* Some correlations between mobility and mobile service usage were observed in the literature [18, 19, 8]. We study whether this occurs with mobile data traffic volume predictability as well. To that end, we compute, for each user, the radius of gyration [31], *i.e.*, the root mean square distance of all recorded locations with respect to their center of mass. The radius of gyration provides a measure of the overall level of mobility of an individual and allows classifying subscribers into the following categories: sedentary, urban, peri-urban, and long-range commuters [32]. Fig. 4(c) presents the CDF of $\Pi^{\max}(\mathbb{V})$ computed on each user category. Again, there exists a slight shift towards lower values in the $\Pi^{\max}(\mathbb{V})$ distribution, as the level of mobility grows. However, the variation is minimal, at 1.3% between commuters and sedentary users. This implies that less users may have slightly more regular data traffic patterns, yet the difference is marginal.

In conclusion, we find no significant correlations between dominant subscribers' features and the predictability of the mobile data traffic volume they generate. In fact, all plots in Fig. 4 indicate that the heterogeneity of $\Pi^{\max}(\mathbb{V})$ across all users is fairly low: the high predictability of traffic volume is a property shared throughout the whole user population.

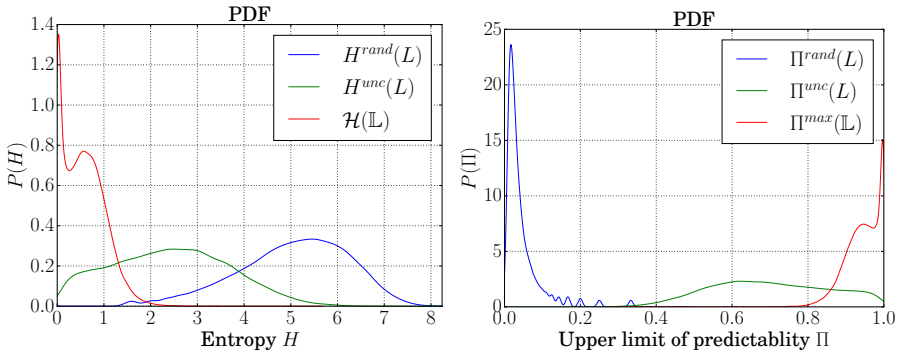
6 Joint predictability of traffic and mobility

In this section, we push our analysis further, and study the joint predictability of future mobile data traffic volume and visited locations, on a per-user basis. In other words, we investigate how predictable is the combination of *how much* traffic is generated by a mobile phone subscriber and *where* this happens. Note that the temporal dimension, *i.e.*, *when* the mobile data traffic is consumed, is implicitly taken into account by the temporal correlation in the definition of the predictability upper bound. Overall, our analysis provides a first comprehensive understanding of whether it is possible to anticipate when, where and how much mobile data traffic is generated by individual subscribers.

6.1 Methodology

We build on knowledge of each user's sequences of traffic volumes $\mathbb{S}_u^{vol}$ and locations $\mathbb{S}_u^{loc}$, and compute several measures of interest, as follows.

Predictability of mobility. We consider a stationary stochastic process $\mathbb{L} = \{L_i, \cdots, L_T\}$ which represents a sequence of locations L_i, recorded for a given user at each time interval i. In a similar way to what done in Sec. 5.1 for the mobile data traffic volume, we leverage $\mathbb{S}_u^{loc}$ to calculate



(a) Entropy (b) Predictability

Figure 5: (a) The distribution of the random entropy $H^{\text{rand}}(L)$, the temporal-uncorrelated entropy $H^{\text{unc}}(L)$ and the entropy rate $\mathcal{H}(\mathbb{V})$ across the observed 45,832 users. (b) The distribution of the upper bounds on the predictability $\Pi^{\text{rand}}(L)$, $\Pi^{\text{unc}}(L)$ and $\Pi^{\text{max}}(\mathbb{L})$ across the observed users.

three entropy rate variants on subscribers' mobility: the *random entropy* $H^{\text{rand}}(L)$, the *temporal-uncorrelated entropy* $H^{\text{unc}}(L)$, and the actual *entropy rate* $\mathcal{H}(\mathbb{L})$. These allow computing their corresponding upper bounds on the predictability $\Pi^{\text{rand}}(L)$, $\Pi^{\text{unc}}(L)$, and $\Pi^{\text{max}}(\mathbb{L})$. These are the same exact measures used in [3], and are used to study to what extent user mobility can be anticipated when considered in isolation. These are the exact same measures used in [3], and are used to study to what extent user mobility can be anticipated when considered in isolation.

Predictability of joint mobility and traffic volume. The traffic process $\mathbb{V}$ and mobility process $\mathbb{L}$ are combined into a single joint process $\mathbb{M} = \{(V_1, L_1), \cdots, (V_T, L_T)\}$. Correspondingly, from a measurement data perspective, $\mathbb{S}_u^{vol}$ and $\mathbb{S}_u^{loc}$ are merged into $\mathbb{S}_u^{mix} = \{(v_u^1, l_u^1), \cdots, (v_u^T, l_u^T)\}$, for each user u. The following variants of the entropy rate are calculated on $\mathbb{S}_u^{mix}$ $\forall u \in \mathcal{U}$. The *temporal-uncorrelated entropy* $H^{\text{unc}}(V, L) \equiv -\sum_{v \in V, l \in L} p(v, l) \log_2 p(v, l)$ determines the heterogeneity deriving from simply considering the user's location and traffic volume together. The *joint actual entropy rate* $\mathcal{H}(\mathbb{V}, \mathbb{L})$ is defined as the actual entropy rate of the joint stationary process $\mathbb{M}$. It expresses the combined uncertainty of a user's location and traffic volume at a given time instant, considering his previous history of movements and mobile service usage. The corresponding predictability upper bounds $\Pi^{\text{unc}}(V, L)$ and $\Pi^{\text{max}}(\mathbb{V}, \mathbb{L})$ are calculated as detailed in Sec. 3.

Predictability of data traffic conditioned on mobility. This is a simplified variant of the joint case above, where the two dimensions of traffic volume and mobility are not considered at once, but the former is conditioned on the latter. In other words, only the traffic volume is forecast, assuming knowledge of the past and current locations. Formally, the *conditional entropy rate* is $\mathcal{H}(\mathbb{V}|\mathbb{L}) \equiv \mathcal{H}(\mathbb{V}, \mathbb{L}) - \mathcal{H}(\mathbb{L})$, and the *temporal-uncorrelated conditional entropy* is $H^{\text{unc}}(V|L) \equiv H^{\text{unc}}(V, L) - H^{\text{unc}}(L)$. The predictability upper bounds mapping to each entropy rate are $\mathbb{S}_u^{loc}$ and $\mathbb{S}_u^{vol}$, respectively.

6.2 User mobility in isolation

We first investigate how predictable is individual mobility when considered in isolation. This means to re-run the exact same study presented in [3] on our CDR dataset. The results, in Fig. 5, are consistent with those in the literature, yet they present interesting slight variations that are discussed next.

Fig. 5(a) presents the PDF of the entropy rates $H^{\text{rand}}(L)$, $H^{\text{unc}}(L)$ and $\mathcal{H}(\mathbb{L})$. $P(H^{\text{rand}}(L))$ and $P(H^{\text{unc}}(L))$ are bell-shaped and have statistical measures that are close to those found in [3]. Specifically, $H^{\text{rand}}(L)$ has a mean of approximately 5 bit, indicating that a user's entire movement space is composed of $2^{H^{\text{rand}}(L)} \approx 32$ cells on average. Also, $H^{\text{rand}}(L)$ has a significant variance, *i.e.*, the geographical span of movements varies widely from person to person in our dataset. The mean is higher in the case of $H^{\text{unc}}(L)$, at ≈ 2.4 bit. Users tend to favor some locations over others, and keeping this into consideration allows making more accurate forecasts

on the next location they will visit: the uncertainty shrinks to just $2^{H^{\text{unc}}(L)} \approx 5$ locations on average.

However, the most interesting result is the distribution of $\mathcal{H}$, which, unlike what found in previous studies, is the composition of two bell-shaped distributions with peaks at around 0.03 and 0.65, respectively. The second peak implies that the uncertainty in anticipating the next cell of a subscriber is limited to $2^{\mathcal{H}(\mathbb{L})} \approx 2$ options; this is consistent with what reported in [3]. The first peak is symptomatic of an even lower indecision: in fact, $2^{0.03} \approx 1$, *i.e.*, there are situations where the next location of a user is almost certain. We ascribe this result to the fact that we complete the original CDR data using the `stop-by-spothome` approach, as discussed in Sec. 4.2.1. On the one hand, the data completion places individuals at one single cell (*i.e.*, their home location) for long, continuative periods of time: there is thus little uncertainty about where subscribers are at night. On the other hand, raw CDR data was employed in [3], which could not capture the low entropy rate associated with the lack of movement overnight.

Fig. 5(b) shows the PDF of the predictability upper bounds $\Pi^{\text{rand}}(L)$ and $\Pi^{\text{unc}}(L)$, corresponding to the entropy rates above. The results are consistent: $\Pi^{\text{rand}}(L)$ has a distribution that is narrowly peaked at very low values, while $\Pi^{\text{unc}}(L)$ yields better predictability but widely varies significantly among individuals. While $\Pi^{\text{rand}}(L)$ distribution is in agreement with those originally presented in [3], our results on the $\Pi^{\text{unc}}(L)$ distribution are instead much better (*i.e.*, $\Pi^{\text{unc}}(L) = 0.6$, instead of 0.3 as reported in [3, Fig. 2.B]), what is due to our CDR dataset completion process. More significant differences emerge instead for $\Pi^{\max}(\mathbb{L})$, with two distinct peaks at $\Pi^{\max} = 0.94$ and $\Pi^{\max} = 0.99$. The former value is very close to the $\Pi^{\max} = 0.93$ reported in [3]. The second peak is again due to the CDR dataset completion process, as explained above[2]. In all cases, our results confirm that user mobility alone is highly predictable.

6.3 Mobile data traffic volume and mobility

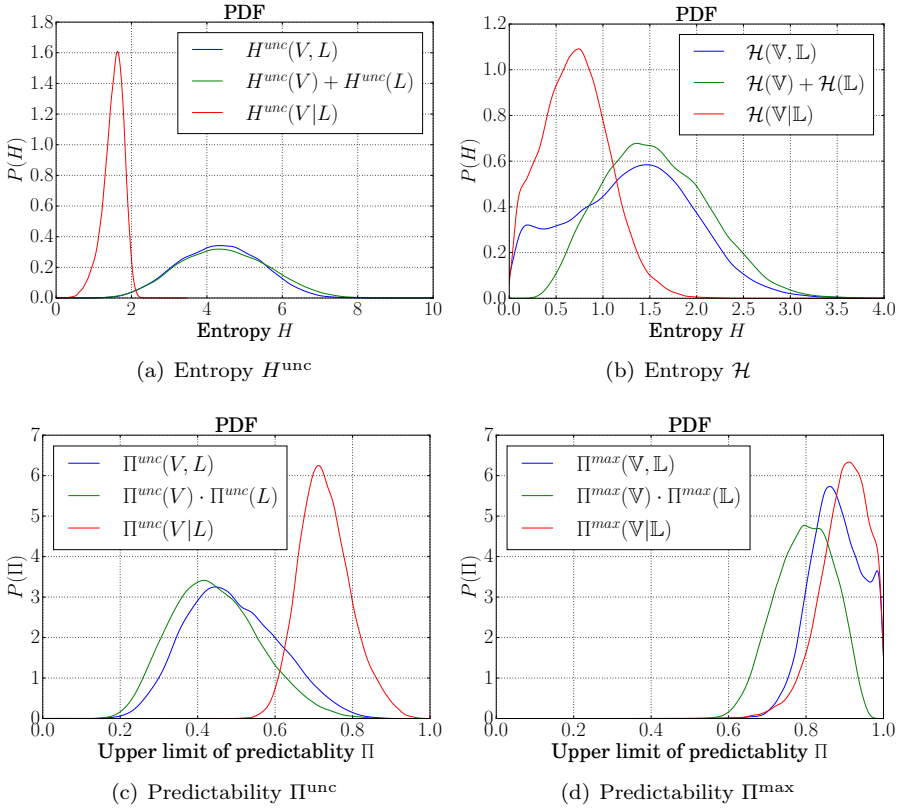
We now consider the uncertainty and predictability of traffic volume and mobility at once. Our results are summarized in Fig. 6, which portrays PDFs for different ways of bringing together the two dimensions of traffic volumes and locations. Specifically, each plot contains three curves: *(i)* the joint entropy or associated predictability, *(ii)* the sum of entropy rates measured for traffic volume and mobility separately, and *(iii)* the conditional entropy rate or associated predictability. The second curve represents the uncertainty (or predictability) in the case the stochastic processes driving mobility and traffic volume consumption are independent of each other. Fig. 6(a) and Fig. 6(c) refer to temporal-uncorrelated versions, whereas Fig. 6(b) and Fig. 6(d) concern our actual measures of interest.

A first interesting remark is that $H^{\text{unc}}(V, L)$ and $H^{\text{unc}}(V) + H^{\text{unc}}(L)$ in Fig. 6(a), and consequently $\Pi^{\text{unc}}(V, L)$ and $\Pi^{\text{unc}}(V) \cdot \Pi^{\text{unc}}(L)$ in Fig. 6(c), are nearly indistinguishable. Instead, $\mathcal{H}(\mathbb{V}, \mathbb{L})$ and $\mathcal{H}(\mathbb{V}) + \mathcal{H}(\mathbb{L})$ in Fig. 6(b), and consequently $\Pi^{\max}(\mathbb{V}, \mathbb{L})$ and $\Pi^{\max}(\mathbb{L}) \cdot \Pi^{\max}(\mathbb{V})$ in Fig. 6(d), show significant differences. Hence, there exists some correlation between the mobility and traffic volume consumption processes, and such correlation mainly emerges when considering – and it is thus driven by – the temporal ordering of events. As observed in Fig. 6(d), a joint prediction of the amount of traffic consumed next, and of the future location where this occurs, can yield a better accuracy than forecasting the two separately, when knowledge of the previous actions of the individual is taken into account. The shift between $\Pi^{\max}(\mathbb{L}) \cdot \Pi^{\max}(\mathbb{V})$ and $\Pi^{\max}(\mathbb{V}, \mathbb{L})$ is of 10% on average.

More importantly, we note that the mean value of $\Pi^{\max}(\mathbb{V}, \mathbb{L})$ is at 0.88, with the probability mass above 0.8 and a noticeable peak at 0.98. Therefore, our main conclusion is that it is possible to anticipate how much mobile data traffic (as an order of magnitude) will be consumed by a given subscriber and where this will occur in a very effective manner (*i.e.*, with an 88% accuracy on average), by knowing the past history of activities of the target individual.

If the available information about each user increases, and the location information can be precisely established (*e.g.*, because mobility occurs at much longer timescales than service consumption, or we know that the user has especially deterministic movement pattern), one can remove the uncertainty about the mobility dimension. In fact, the knowledge about the past and current locations can be then leveraged to even improve the accuracy of the prediction,

[2]We verified this hypothesis by running experiments on our CDR dataset without data completion: in this case, we obtain a bell-shaped distribution of $\Pi^{\max}(\mathbb{L})$ peaked at 0.93, and the second peak disappears.



(a) Entropy H^{unc} (b) Entropy $\mathcal{H}$

(c) Predictability Π^{unc} (d) Predictability Π^{max}

Figure 6: (a) Distributions of the different flavors of temporal-uncorrelated entropies: $H^{\text{unc}}(V,L)$, $H^{\text{unc}}(V)+H^{\text{unc}}(L)$ and $H^{\text{unc}}(V|L)$. (b) Distributions of the different flavors of entropy rates: $\mathcal{H}(\mathbb{V})$, $\mathcal{H}(\mathbb{V})+\mathcal{H}(\mathbb{V})$ and $\mathcal{H}(\mathbb{V}|\mathbb{L})$. (c) Distributions of the predictability upper bounds $\Pi^{\text{unc}}(V,L)$, $\Pi^{\text{unc}}(L)\cdot\Pi^{\text{unc}}(L)$ and $\Pi^{\text{unc}}(V|L)$ based on the corresponding temporal-uncorrelated entropies. (d) Distributions of the predictability upper bounds $\Pi^{\text{max}}(\mathbb{V},\mathbb{L})$, $\Pi^{\text{max}}(\mathbb{V})\cdot\Pi^{\text{max}}(\mathbb{L})$ and $\Pi^{\text{max}}(\mathbb{V}|\mathbb{L})$ based on the corresponding entropy rates.

which occurs in both temporal-uncorrelated and actual cases, as shown in Fig. 6. The plot in Fig. 6(d) also portrays the range of the predictability gain: by comparing $\Pi^{\text{max}}(\mathbb{V}|\mathbb{L})$ with $\Pi^{\text{max}}(\mathbb{V})$ in Fig. 2(b), we observe that including location information in the prediction process allows forecasting the future consumption of mobile data traffic with 5% higher accuracy, pushing the overall performance from 85% to 90%. Hence, our second conclusion is that using knowledge of the spatio-temporal trajectories of subscribers can further improve the design of a prediction model targeting individual traffic volume consumption. Yet, the gain is not dramatic with respect to a technique that only relies on temporal information.

7 Conclusions

In this paper, we have explored the predictability of individual mobility and mobile data traffic consumption. To the best of our knowledge, this is the first work that jointly considers user's location and mobile service usage in a per-user predictability analysis. We found an upper limit on such predictability, demonstrating that it is possible to anticipate how much traffic a subscriber will generate, as well as where he will do so, with 88% accuracy on average, by leveraging historical information about the user. This result is possible thanks to correlations between visited locations and traffic volumes: Indeed, trying to predict traffic volumes in isolation reduces the accuracy to 85%. If the location information is instead known and used to predict the future volume, the mean precision grows to 90%.

References

[1] W. Su, S.-J. Lee, and M. Gerla, "Mobility prediction in wireless networks," in *MILCOM 2000. 21st Century Military Communications Conference Proceedings*, vol. 1. IEEE, 2000, pp. 491–495.

[2] P. N. Pathirana, A. V. Savkin, and S. Jha, "Mobility modelling and trajectory prediction for cellular networks with mobile base stations," in *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*. ACM, 2003, pp. 213–221.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[4] J. Yao, S. S. Kanhere, and M. Hassan, "An empirical study of bandwidth predictability in mobile computing," in *Proceedings of the third ACM international workshop on Wireless network testbeds, experimental evaluation and characterization*. ACM, 2008, pp. 11–18.

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

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

[8] 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.

[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*. IEEE, 2015, pp. 230–235.

[10] 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*. ACM, 2013, pp. 321–326.

[11] 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.

[12] A. Sang and S.-q. Li, "A predictability analysis of network traffic," *Computer networks*, vol. 39, no. 4, pp. 329–345, 2002.

[13] K. Shah, S. Bohacek, and E. A. Jonckheere, "On the predictability of data network traffic," in *Proceedings of the American Control Conference*, vol. 2, 2003, pp. 1619–1624.

[14] Y. Baryshnikov, E. Coffman, G. Pierre, D. Rubenstein, M. Squillante, and T. Yimwadsana, "Predictability of web-server traffic congestion," in *10th International Workshop on Web Content Caching and Distribution (WCW'05)*. IEEE, 2005, pp. 97–103.

[15] 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.

[16] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 1, Jun. 2011.

[17] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," Ph.D. dissertation, INRIA, 2014.

[18] 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.

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

[20] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[21] T. Schürmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 6, no. 3, pp. 414–427, Sep. 1996.

[22] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1319–1327, 1998.

[23] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. New York, New York, USA: ACM, Sep. 2007, pp. 123–134.

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

[25] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the Limit of Predictability in Human Mobility," *Scientific reports*, vol. 3, p. 2923, Oct. 2013.

[26] 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.

[27] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "NextPlace: A Spatio-temporal Prediction Framework for Pervasive Systems." *Pervasive*, vol. 6696, no. Chapter 10, pp. 152–169, 2011.

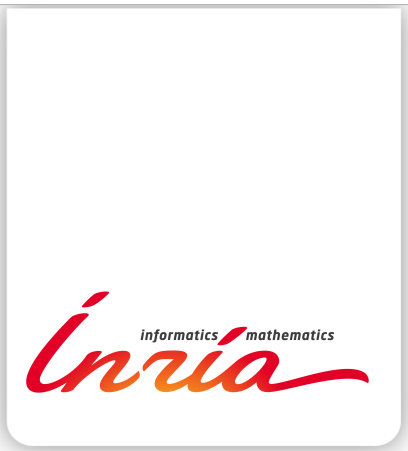
[28] A.-l. Alcatel-Lucent, "9900 wireless network guardian," *White Paper, Dec*, 2012.

[29] S. Hoteit, G. Chen, A. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis," in *ACM Chants*, 2016.

[30] G. Smith, R. Wieser, J. Goulding, and D. Barrack, "A refined limit on the predictability of human mobility," in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, March 2014, pp. 88–94.

[31] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[32] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, 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