2017 IEEE 42nd Conference on Local Computer Networks

The Spatiotemporal Interplay of Regularity and Randomness in Cellular Data Traffic

Guangshuo Chen*†, Sahar Hoteit‡, Aline Carneiro Viana*, Marco Fiore§ and Carlos Sarraute¶
*École Polytechnique, Université Paris Saclay, France, guangshuo.chen@inria.fr
†INRIA, Université Paris Saclay, France, aline.viana@inria.fr
‡ Laboratoire des Signaux et Systèmes, Université Paris Sud-CNRS-CentraleSupélec,
Université Paris-Saclay, France, sahar.hoteit@u-psud.fr
§CNR - IEIIT, Italy, marco.fiore@ieiit.cnr.it
¶Grandata Labs, USA, charles@grandata.com

***Abstract*—In this paper, we leverage two large-scale real-world datasets to provide the first results on the limits of predictability of cellular data traffic demands generated by individual users over time and space. Using information theory tools, we measure the maximum predictability that any algorithm has potential to achieve. We first focus on the predictability of mobile traffic consumption patterns in isolation. Our results show that it is theoretically possible to anticipate the individual demand with a typical accuracy of 85% and reveal that this percentage is consistent across all user types. 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.**

***Index Terms*—Fundamental limits; user mobility; user data traffic; call detail records; performance analysis.**

I. Introduction

The quantitative understanding of human behavior (*e.g.*, user's whereabouts or data traffic) has recently emerged as a central question in multi-disciplinary research [1]. The performance of any practical technique that aims at anticipating human behaviors is bounded by *predictability*, measuring to what degree a specific behavior can be foreseen. In the area of wireless communication, the predictability bound of mobile data traffic predictors opens new opportunities to network operators to *(i)* manage their resources in advance, *(ii)* accommodate the future demand at lower maintenance and operational costs, and *(iii)* define traffic plans that are better tailored to users' needs. Also, the predictability bound contributes to the understanding of human nature as well as the design of effective prediction algorithms.

In this paper, we analyze the traffic predictability from the viewpoint of individual users instead of base stations or aggregated perspective [2], [3]. For mobile phone users, the regularity of their Internet traffic, which drives the predictability together with randomness, is not only reflected by statistical measures of their consumed data volumes [4] but also by their whereabouts [5]. However, to the best of our knowledge, there is no analysis of *(i)* how per-user regularity of mobile data traffic is translated into actual predictability, or *(ii)* the associated impacts brought by jointly considering users' visited locations.

We aim at filling the gap above, and provide a first investigation of the predictability of mobile data traffic generated by individual users. Specifically, we focus on data traffic volume, *i.e.*, the amount of bytes generated by the mobile services consumed by a given user. Overall, our study allows answering an important question: *to what degree is the individual consumption of mobile data traffic predictable?*

We address this problem by studying the variation in individual's mobile data traffic over time and space and investigating its predictability limits by using tools from information theory. In summary, our contributions are:

- We provide a first study of the predictability of mobile data traffic usage from the viewpoint of individual users. Our work is powered by two large-scale real-world datasets, presented in Sec. II.
- We derive a promising upper bound (85%) to the performance of practical algorithms for the prediction of the volume of mobile data traffic from the user's historical usage. Details are provided in Sec. III.
- We also introduce a first investigation on the power of jointly forecasting when, where, and how much mobile data traffic is generated by individual users. We observe that the strong correlation between mobility and mobile service usage leads to a performance gain on the upper predictability bound from 85% to 90%. Details are provided in Sec. IV.
- Our discussion about the cause of the high (joint) predictability sheds light on the design of predicting algorithms. Details are provided in Sec. V.

II. Data overview

Our study is based on the observation across the population of approximately 45K mobile phone users during a consecutive period of 92 days in 2014. Footprints and Internet data traffic demands are provided by two anonymous datasets having the same hashed users' identifiers, respectively:

1) CDR (Call Detail Record) dataset: A CDR entry is logged once a user initiates or receives a voice call [1]. Each CDR entry contains the user's identifier, the call start time and the handling cell tower location. The users and their CDRs are collected from a major cellular operator in a metropolitan area where a cell tower covers around 2 km^2 on average, leading to a fair spatial granularity in localizing the users.

2) Internet data session dataset: This dataset describes how the users generate Internet data traffic through their devices and consists of *Internet data sessions*. Every data session is

DOI 10.1109/LCN.2017.41

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 session entry contains the user's identifier, the volume of upload and download data (in KiloBytes) and the session start time.

Those $45K$ users meet the followng criteria: *(i)* they have visited at least 2 locations; *(ii)* they have CDR-based footprints in at least 20% of the observing hours; *(iii)* they establish sessions in at least 73 days (*i.e.*, 80% of the observing period). The criteria ensures statistical significance of our analysis.

Given the two datasets, we compute representative discretized time series of data traffic volumes $\mathbb{S}_u^{vol}$ and of locations $\mathbb{S}_u^{loc}$ per user as follows:

1) Volume processing: The Internet session dataset allows processing sessions and quantizing volume in a straightforward way. Hence, for a user $u \in \mathcal{U}$ (where $\mathcal{U}$ is the user set), her volume time series, represented as $\mathbb{S}_u^{vol} = \{v_u^1, v_u^2, \ldots, v_u^i, \ldots, v_u^T\}$, is computed by aggregating the volume of traffic from all her sessions recorded during each one-hour interval. Particularly, the observing period is divided into $T = 24$ hours/day $\times$ 92 days $= 2208$ intervals, and v_u^i describes the usage of data traffic of the user u in the i^{th} interval and is symbolized as one of the eight quantization intervals: 0 (*idle*), $(1, 10)$, $(10, 10^2)$, $\ldots$, $(10^6, 10^7)$ in KB.

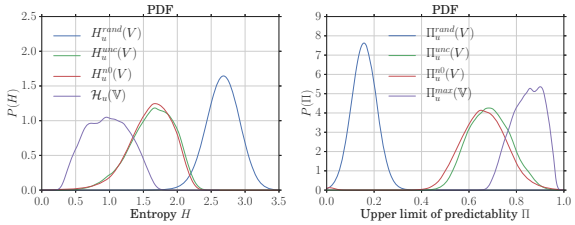
2) Location processing: An imputation is necessary before processing CDRs because the mobility information is actually incomplete. For that, we firstly apply the `stop-by-spothome` approach, a CDR completion approach introduced in [6]. This significantly increases the temporal coverage of CDRs, particularly overnight, without affecting the localization precision. Then for a user u, his movement is represented by a time series of locations as $\mathbb{S}_u^{loc} = \{\ell_u^1, \ell_u^2, \ldots, \ell_u^i, \ldots, \ell_u^T\}$, where ℓ_u^i is the location of u in the i^{th} one-hour interval. On each interval, ℓ_u^i is the cell tower location where the user u spend the most time or *unknown* if the user has no calls during the interval. $\mathbb{S}_u^{loc}$ and $\mathbb{S}_u^{vol}$ of each user cover the same observing period and thus have the same length. For more details on the data and preprocessing, we refer the reader to the full paper [7].

III. Predictability of Mobile Data Traffic

We first study the predictability of the mobile data traffic generated by individual subscribers by focusing on the forecast of traffic volume in isolation.

Methodology: For each user, we consider her $\mathbb{S}_u^{vol}$ as a consecutive sequence of sampled from a corresponding model describing how she generates data traffic. Given the model, we then estimate the entropy rate as $H_u(V) = \lim_{t\to\infty} H(V_u^t | V_u^{t-1}, \cdots, V_u^1)$ from the $\mathbb{S}_u^{vol}$ as a prepositive step of measuring the predictability. The entropy rate measures the average uncertainty of anticipating data traffic quantization as $V_u^{(\cdot)}$ on each interval. Hereby, we derive four variants of the entropy rate from different models, as follows:

- The *random entropy* rate is formulated as $H_u^{\text{rand}}(V) \equiv \log_2 N$, where N is the number of distinct quantizations of traffic volumes, derived from an equally probable and time-independent model.
- The *temporal-uncorrelated entropy* rate is formulated as $H_u^{\text{unc}}(V) \equiv -\sum_{v\in V} P(v)\log_2 P(v)$ where the user's traffic follows a heterogeneous and time-independent model. This entropy rate characterizes the heterogeneity of a mobile demand model that has no temporal correlations, hence its name.
- The *nonzero-temporal-uncorrelated entropy* rate is based on the same model of $H_u^{\text{unc}}(V)$, but it is limited to those cases when the user is not idle. Formally, it is $H_u^{\text{n0}}(V) \equiv -\sum_{v\in V/\{0\}} P(v|v \neq 0)\log_2 P(v|v \neq 0)$. It captures the heterogeneity of the traffic volume exchanged during active hours only, yet still ignoring temporal correlations.
- Finally, the *actual entropy* rate, denoted by $\mathcal{H}_u(\mathbb{V})$, is estimated based on Lempel-Ziv compression [8], where the data traffic is considered to follow a stochastic process without other constraints. This entropy rate depends not only on the frequency of appearance of each discretized traffic volume but also on the order in which they appear, capturing the temporal order presented in a subscriber's traffic usage pattern.



(a) Entropy (b) Predictability

Fig. 1. (a) Distributions of the random entropy $H_u^{\text{rand}}(V)$, the temporal-uncorrelated entropy $H_u^{\text{unc}}(V)$, the nonzero-temporal-uncorrelated entropy $H_u^{\text{n0}}(V)$, and the entropy rate $\mathcal{H}_u(\mathbb{V})$ as observed in the individual traffic demand. (b) Equivalent distributions of the upper bounds on the predictability $\Pi_u^{\text{rand}}(V)$, $\Pi_u^{\text{unc}}(V)$, $\Pi_u^{\text{n0}}(V)$ and $\Pi_u^{\text{max}}(\mathbb{V})$.

The entropy rate and predictability are negatively correlated variables: a behavior with low (or high) uncertainty is highly (or little) predictable. Mathematically, given an entropy rate variant H, its predictability Π satisfies $\Pi \leq \Phi^{-1}(H)$ where $\Phi(x) \equiv (1-x)log(N-1) - x\log x - (1-x)\log(1-x)$ and $\Phi^{-1}(x)$ is its inverse function [9]. In our context, the upper bound of Π is an estimation of the maximum achievable accuracy in the prediction of the mobile traffic demand, given a particular model. Hence, four upper bounds for the predictability, *i.e.*., $\Pi_u^{\text{max}}(\mathbb{V})$, $\Pi_u^{\text{rand}}(V)$, $\Pi_u^{\text{unc}}(V)$, and $\Pi_u^{\text{n0}}(V)$, are calculated from the entropy rates.

Results: Let us start by considering the PDF of $H_u^{\text{rand}}(V)$ in Fig. 1(a). It peaks at $H_u^{\text{rand}}(V) \approx 2.7$ which indicates that on average, the quantization level of traffic volume can be represented by at most three bits per time interval of new information; that is, a user who randomly demand traffic can generate traffic on average in $2^{H_u^{\text{rand}}(V)} = 2^3 \approx 8$ quantization levels. This phenomenon is normal, as most of users have eight traffic volume quantization levels (up to 10GB/hr). When the temporal-uncorrelated entropy, i.e., $H_u^{\text{unc}}(V)$, is concerned, a sizable shift of probability occurs: the uncertainty decreases to $2^{H_u^{\text{unc}}(V)} = 2^{1.63} \approx 3$. Under this model, each user tends to generate traffic that is described by just three quantization levels out of the eight available.

Interestingly, idle time intervals do not bias such regularity.

Indeed, the PDF of $H_u^{\text{n0}}(V)$ overlaps well to that of $H_u^{\text{unc}}(V)$, 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 $\mathcal{H}_u(\mathbb{V})$, which peaks at 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. 1(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 $\Pi_u^{\text{rand}}(V)$ peaks at 0.16, i.e., it is very hard to guess the volume of traffic generated by such a stochastic model. The predictability grows for $\Pi_u^{\text{unc}}(V)$ and $\Pi_u^{\text{n0}}(V)$, which peak at 0.69 and 0.66, respectively. More importantly, $\Pi_u^{\text{max}}(\mathbb{V})$ 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 *the traffic volume which subscribers generate via their mobile devices is highly predictable*.

IV. Joint Predictability of Traffic and Mobility

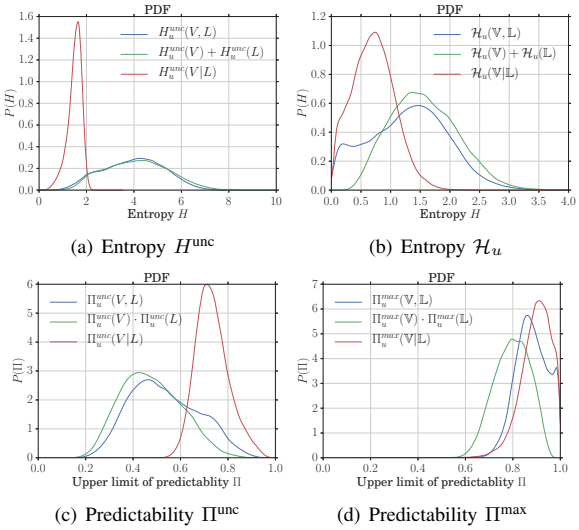
We further study the joint predictability of future mobile data traffic volume and visited locations, on a per-user basis. We investigate how predictable is the combination of *how much* traffic is generated by a mobile phone user and *where* this happens. Note that the temporal dimension, *i.e.*, *when* the mobile data traffic is consumed, is implicitly taken into account by the temporal correlation in the definition of the predictability upper bound.

Methodology: Firstly, as in Sec. III, we leverage $\mathbb{S}_u^{loc}$ to calculate two entropy rate variants on subscribers' mobility: the *temporal-uncorrelated entropy* rate $H_u^{\text{unc}}(L)$ and the actual *entropy rate* $\mathcal{H}_u(\mathbb{L})$ and their corresponding upper bounds on the predictability $\Pi_u^{\text{unc}}(L)$ and $\Pi_u^{\text{max}}(\mathbb{L})$.

Secondly, from a user's $\mathbb{S}_u^{loc}$ and $\mathbb{S}_u^{vol}$, we compute two joint entropy variants that consider the user's volume and location together. The *temporal-uncorrelated entropy* rate $H_u^{\text{unc}}(V,L) \equiv -\sum_{v\in V, l\in L} P(v,l)\log_2 P(v,l)$ determines the joint heterogeneity of the user's location and traffic volume. The *joint actual entropy rate* $\mathcal{H}_u(\mathbb{V},\mathbb{L})$ is defined as the actual entropy rate of the joint stationary process that generates $\mathbb{S}_u^{vol}$ and $\mathbb{S}_u^{loc}$. 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. Also, the corresponding predictability upper bounds $\Pi_u^{\text{unc}}(V,L)$ and $\Pi_u^{\text{max}}(\mathbb{V},\mathbb{L})$ are calculated.

We also measure the *conditional entropy rate* as $\mathcal{H}_u(\mathbb{V}|\mathbb{L}) \equiv \mathcal{H}_u(\mathbb{V},\mathbb{L}) - \mathcal{H}_u(\mathbb{L})$ and the *temporal-uncorrelated conditional entropy* rate as $H_u^{\text{unc}}(V|L) \equiv H_u^{\text{unc}}(V,L) - H_u^{\text{unc}}(L)$ along with their corresponding predictability upper bounds, respectively. The conditional variants corresponds to the case that only the traffic volume is forecast, assuming knowledge of the past and current locations.

Results: Fig. 2 summarizes results with respect to the uncertainty and predictability of traffic volume and mobility, for different ways of bringing together the two dimensions of traffic volumes and locations. Fig. 2(a) and Fig. 2(c) refer to temporal-uncorrelated versions, whereas Fig. 2(b) and Fig. 2(d) concern our actual measures of interest.



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

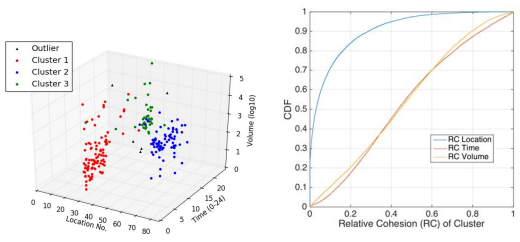
(c) Predictability Π^{unc} (d) Predictability Π^{max}

Fig. 2. (a) Distributions of the different flavors of temporal-uncorrelated entropy variants: $H_u^{\text{unc}}(V,L)$, $H_u^{\text{unc}}(V)+H_u^{\text{unc}}(L)$ and $H_u^{\text{unc}}(V|L)$. (b) Distributions of the different flavors of entropy rates: $\mathcal{H}_u(\mathbb{V})$, $\mathcal{H}_u(\mathbb{V})+\mathcal{H}_u(\mathbb{V})$ and $\mathcal{H}_u(\mathbb{V}|\mathbb{L})$. (c) Distributions of the predictability upper bounds $\Pi_u^{\text{unc}}(V,L)$, $\Pi_u^{\text{unc}}(L)\cdot\Pi_u^{\text{unc}}(L)$ and $\Pi_u^{\text{unc}}(V|L)$ based on the corresponding temporal-uncorrelated entropies. (d) Distributions of the predictability upper bounds $\Pi_u^{\text{max}}(\mathbb{V},\mathbb{L})$, $\Pi_u^{\text{max}}(\mathbb{V})\cdot\Pi_u^{\text{max}}(\mathbb{L})$ and $\Pi_u^{\text{max}}(\mathbb{V}|\mathbb{L})$ based on the corresponding entropy rates.

A first interesting remark is that $H_u^{\text{unc}}(V,L)$ and $H_u^{\text{unc}}(V)+H_u^{\text{unc}}(L)$ in Fig. 2(a), and consequently $\Pi_u^{\text{unc}}(V,L)$ and $\Pi_u^{\text{unc}}(V)\cdot\Pi_u^{\text{unc}}(L)$ in Fig. 2(c), are nearly indistinguishable. Instead, $\mathcal{H}_u(\mathbb{V},\mathbb{L})$ and $\mathcal{H}_u(\mathbb{V})+\mathcal{H}_u(\mathbb{L})$ in Fig. 2(b), and consequently $\Pi_u^{\text{max}}(\mathbb{V},\mathbb{L})$ and $\Pi_u^{\text{max}}(\mathbb{L})\cdot\Pi_u^{\text{max}}(\mathbb{V})$ in Fig. 2(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. 2(d), a joint prediction of the next consumed amount of traffic 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 $\Pi_u^{\text{max}}(\mathbb{L})\cdot\Pi_u^{\text{max}}(\mathbb{V})$ and $\Pi_u^{\text{max}}(\mathbb{V},\mathbb{L})$ is of 10% on average.

More importantly, we note that the mean value of $\Pi_u^{\text{max}}(\mathbb{V},\mathbb{L})$ 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 user 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, one can remove the uncertainty about the mobility dimension. This would improve the accuracy of the prediction, which occurs in both temporal-uncorrelated and actual cases, as shown in Fig. 2. The plot in Fig. 2(d) also portrays the range of the



(a) Example: User 1 (b) Relative Cohesion

Fig. 3. (a) An Example of mapping a user's sessions mapped the three-dimensional space. (b) CDF of each cluster's relative cohesion on the three dimensions. Figure best viewed in colors.

predictability gain: by comparing $\Pi_u^{\max}(\mathbb{V}|\mathbb{L})$ with $\Pi_u^{\max}(\mathbb{V})$ in Fig. 1(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 users 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.*

V. Discussion and Conclusion

Discussion: In order to understand the cause of the high (joint) predictability observed in Sec. IV, we map each user's Internet sessions into a three-dimensional space of *location*, *time*, and *volume*, so as to have an intuitive spatiotemporal representation. Each session becomes then a point $p(l,t,v)$ into this space. Note that we express l as the linear ordering of the corresponding bidimentional locations, as returned by Optics [10]: a density-based cluster algorithm that places spatially close bidimentional locations as neighbors in the ordering l. Time t is expressed by hours with decimals from 0 to 24, where the date is ignored. Volume v is the magnitude of the traffic volume, *i.e.*, $log_{10}(\cdot)$. This mapping reveal how a user generates mobile Internet sessions. In Fig. 3(a), most sessions are aggregated on two major locations (30 and 60), probably mapping to home and working place according to their time of visits. Sessions containing data traffic over 10MB mainly occur at the location 30 during nighttime.

To quantitatively investigate the 3D space representation of $p(l,t,v)$ points, we use DBScan [11] to cluster each user's sessions in the three-dimensional space. For the clustering, a weighted euclidean distance is measured between every two points $p_1(l_1,t_1,v_1)$ and $p_2(l_2,t_2,v_2)$, where the distance of each dimension is computed as follows: *(i)* $dist^{(location)}(p_1,p_2) = \omega_l|\mathbf{l_1} - \mathbf{l_2}|_{geo}$ in kilometers; *(ii)* $dist^{(time)}(p_1,p_2) = \omega_t|t_1 - t_2|$ in hours; *(iii)* $dist^{(volume)}(p_1,p_2) = \omega_v|v_1 - v_2|$ ($|log_{10}\frac{Vol_1}{Vol_2}|$). Each distance is applied to 99% percentile normalization. For each cluster shown in Fig. 3(a), we then use the *relative cohesion* (RC) to quantify the contribution of each dimension to a given cluster as $RC^{(*)} = \frac{\sum_{p\in C} dist^{(*)}(p,c)^2}{\sum_{p\in C} dist(p,c)^2}$, where C and c respectively represent the cluster and its centroid $c = (l_{centroid}, t_{mean}, v_{mean})$. The RCs of the three dimensions satisfy $RC^{(loc)} + RC^{(time)} + RC^{(vol)} = 1$, where $0 < RC^{(*)} < 1$. Hence, if a cluster's RC in one dimension is significantly smaller than the other two dimensions, this dimension is contributing the most to creating the cluster.

Fig. 3(b) shows the distributions of RCs along the three dimensions for all users. The most striking behavior is the much lower RC in space than time or traffic volume: *i.e.*, *"where a user is"* drives the creation of the majority of clusters: The location of a mobile user has a high probability to trigger some routine service consumption activity. Hence, anticipating the future location of a user should be the first target of a solution aiming at predicting mobile user activity. However, we also observe that locations alone do not explain all clusters. A non-negligible fraction of clusters showing high RC in space and low RC in time and traffic volume are also present in several cases. We conclude that the three dimensions are complementary, and though different weights, they are all important for an accurate prediction of the behavior of mobile users. This is consistent with – and explains – our results on the high joint predictability of temporally correlated visited locations and consumed traffic.

Conclusions: To the best of our knowledge, this is the first work to *(i)* show the maximum predictability of *personal* mobile data traffic volume and *(ii)* jointly consider user's location and mobile service usage in a per-user predictability analysis. Our results indicate that there is a large space for predicting mobile data traffic and adapting network optimizing solutions based on the latter, such as for load balancing.

Acknowledgement

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

References

[1] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale Mobile Traffic Analysis: a Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 124–161, 2016.

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

[3] 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.

[4] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," *ACM SIGMETRICS*, vol. 39, June 2011.

[5] 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.

[6] 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.

[7] 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.

[8] T. Schürmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 6, pp. 414–427, Sept. 1996.

[9] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, 1994.

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

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