



HAL
open science

Measurement-driven mobile data traffic modeling in a large metropolitan area

Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Kolar Purushothama Naveen, Carlos Sarraute

► **To cite this version:**

Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Kolar Purushothama Naveen, Carlos Sarraute. Measurement-driven mobile data traffic modeling in a large metropolitan area. [Research Report] RR-8613, 2014, pp.20. hal-01073129v1

HAL Id: hal-01073129

<https://inria.hal.science/hal-01073129v1>

Submitted on 9 Oct 2014 (v1), last revised 16 Jun 2015 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Measurement-driven mobile data traffic modeling in a large metropolitan area

Eduardo Mucelli Rezende Oliveira , Aline Carneiro Viana , K. P.
Naveen, Carlos Sarraute

**RESEARCH
REPORT**

N° 8613

October 2014

Project-Team Infine



Measurement-driven mobile data traffic modeling in a large metropolitan area

Eduardo Mucelli Rezende Oliveira^{*}, Aline Carneiro Viana^{*}, K. P. Naveen, Carlos Sarraute[†]

Équipe-Projet Infine

Rapport de recherche n° 8613 — October 2014 — 20 pages

Résumé : Comprendre la demande de trafic de données mobiles est essentielle pour l'évaluation des stratégies portant sur le problème de l'utilisation de bande passante élevée et l'évolutivité des ressources du réseau, apporté par l'ère "pervasive". Dans cet article, nous effectuons la première modélisation détaillée de l'utilisation du trafic mobile des smartphones dans un scénario métropolitain. Nous utilisons un ensemble de données à grande échelle recueillis au coeur d'un des majeurs réseaux 3G de la capitale du Mexique. Nous analysons d'abord le comportement individuel routinier et nous avons observé des patrons d'utilisation identiques pour les différents jours. Cela nous motive à choisir un jour pour étudier le mode d'utilisation des abonnés (c'est à dire, "quand" et "combien" de trafic est généré) en détail. Nous classons ensuite les abonnés en quatre profils distincts en fonction de leur mode d'utilisation. Nous modélisons enfin le mode d'utilisation de ces quatre profils d'abonnés selon deux périodes différents: de pointe et les heures creuses. Nous montrons que la trace synthétique produite par le modèle de trafic de données imite fidèlement les différents profils d'abonnés en deux périodes, par rapport à l'ensemble de données d'origine.

Mots-clés : réseaux, modèle de trafic, routine

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

[†] 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

Measurement-driven mobile data traffic modeling in a large metropolitan area

Abstract: Understanding mobile data traffic demands is crucial to the evaluation of strategies addressing the problem of high bandwidth usage and scalability of network resources, brought by the pervasive era. In this paper, we conduct the first detailed measurement-driven modeling of smartphone subscribers' mobile traffic usage in a metropolitan scenario. We use a large-scale dataset collected inside the core of a major 3G network of Mexico's capital. We first analyse individual subscribers' routine behaviour and observe identical usage patterns on different days. This motivates us to choose one day for studying the subscribers' usage pattern (i.e., "when" and "how much" traffic is generated) in detail. We then classify the subscribers in four distinct profiles according to their usage pattern. We finally model the usage pattern of these four subscriber profiles according to two different journey periods: peak and non-peak hours. We show that the synthetic trace generated by our data traffic model consistently imitates different subscriber profiles in two journey periods, when compared to the original dataset.

Key-words: networks, traffic model, routine

1 Introduction

The recent boost up of mobile data consumption [1] led by the pervasive era are struggling the 3G cellular networks, which are not always prepared to receive such demand. The steady growth of smartphones, the very rapid evolution of services and their usage is accentuated in metropolitan scenarios due to the high urbanization and concentration of mobile users. Emerging pervasive communication systems will face a number of challenges in this context, including the need to operate in such extreme environments.

In this context, the *understanding of mobile data traffic demands per user* is crucial to the evaluation of data offloading solutions designed to alleviate cellular networks [2, 3], favoring then the proliferation of pervasive communication. Moreover, the definition of a *usage pattern* can allow telecommunication operators to better timely plan network resources allocation and better set subscription plans.

The pervasive era also brought new facilities : Currently smartphones provide the best means of gathering users information about content consumption behaviour on a large scale. In this context, the literature is rich in work studying and modeling users mobility, but little is publicly known about users content consumption patterns.

Contrarily to most related work in the literature based on Call Detail Records (CDRs), we characterize and model real mobile data traffic demands generated by smartphone subscribers. Although convenient and of frequent consideration, CDRs only provide an intuition of users activity in the network : voice calls and SMS. Since smartphones are now used more for data than for calls [4], the use of CDRs for investigating traffic demands is not enough for dimensioning network usages. In addition, besides being sparse in time [5], CDRs do not describe the background traffic load automatically generated by current smartphone applications (e.g., email checks, synchronization).

Instead, our analyses provide a *precise characterization of individual subscribers traffic behaviour clustered by their usage pattern*, instead of a network-wide data traffic view [6, 7, 8]. Note that the high variance of individual subscriber behaviour (in terms of traffic demands and in time) and the use of large scale datasets make this task complex. Our focus in this paper is *to analyse urban mobile data traffic usage pattern* and *to provide a way for synthetically still consistently reproducing it* : The first work in the literature, to the best of our knowledge. The implications of this work are diverse, in particular, in resource allocation planning and testing, or hotspot deployment.

Our study is performed on an anonymized dataset collected at the core of a major 3G network of Mexico's capital (Section 3). The dataset spans 4 months from July to October 2013 and consists of all data traffic associated with 6.8 million subscribers. The dataset describes detailed information on the volume and frequency of any data traffic generated by smartphone subscribers. This includes any uploaded and downloaded data traffic, i.e., not only browsing or messages traffic, but traffic automatically generated by applications are also included. This represents an order of hundreds of Pebibytes (2^{50}) exchanged in the biggest city in Mexico.

We focus on the temporal dynamics of individual subscriber's usage pattern. Thus, we first analyse their traffic usage habits as a function of time (Section 3). We observe identical usage patterns on different days. This motivates us to choose one day for studying the subscribers' usage pattern (i.e., "when" and "how much" traffic is generated) in detail. We then classify the subscribers in four distinct profiles according to their usage pattern (Section 4). We finally model the usage pattern of these four subscriber profiles according to two different journey periods : peak and non-peak hours. Using a test set and numerous statistic tools, we show the effectiveness of our traffic modelling, which is capable of consistently imitating different subscribers profiles in two journey periods, when compared to the original traffic dataset (Section 5).

In summary, our main outcome is a *synthetic measurement-based mobile data traffic generator capable of imitating traffic-related activity patterns of four different categories of subscribers during two time periods of a routinary normal day in their lives*. In this paper, user and subscriber will be used interchangeably.

2 Related Work

The understanding of users' content consumption has attracted significant attention of the networking community in the literature. Its improved understanding is of fundamental importance when looking for *solutions to manage the increased data usage and to improve the quality of communication service provided*. The resulting knowledge can help to design more adaptable networking protocols or services, as well as to determine, for instance, where to deploy networking infrastructure, how to reduce traffic congestion, or how to fill the gap between the capacity granted by the infrastructure technology and the traffic load generated by mobile users.

Most studies in the literature carried out on the analysis of network utilisation are based on the analysis of calling patterns, i.e., Call Detail Records (CDRs). CDRs analysis may provide an idea of the activity of mobile network customers but do not describe realistic data traffic demand patterns [5, 9, 8, 6]. In fact, contrarily to data traffic demands, CDRs have the limitations of being sparse in time (i.e., generated only when a voice call or a short message service occurs), which makes cellular users invisible at all other times. Moreover, due to the richness of the data set used in our studies, we can precisely infer traffic activity patterns over time, instead of only considering the time users actively generated traffic. This includes the traffic load automatically generated by current smartphone applications (email checks, synchronization, etc). Our analysis also differs from [6, 10], since we target an individual user characterization rather than network-wide one. Moreover, contrarily to [8], we focus on an activity pattern characterization of a normal day, known to represent a typical and usual network usage.

Other works that categorized mobile traffic usage [7, 3, 11, 12, 13, 14]. [7, 3, 11, 12] have only considered total traffic volume when characterizing users' behavior. Studying only this metric does not reflect the activity variation of users : i.e., number and frequency of requests. [13, 14] study the distribution of mobile traffic volume among different areas in a specific region. Their study, however, is based on the normalized volume with respect to the total traffic volume in the region. In this work, we do not account for geographic variations of network usages. Instead, we provide a precise network usage characterization of a routinary day of users' life. In this context, users behaviors over time are individually analyzed, with no normalization performed. Similarly, activity patterns and a classification of individual users behavior are considered.

3 Dataset

The final goal of our work is a measurement-driven traffic modeling. The traffic modeling is performed after several measurement-driven analysis of an anonymized dataset provided by a major cellular operator in Mexico. This dataset captures subscribers' traffic activities generated by 6.8 million smartphone devices located within the large urban area of Mexico city. The data includes information about subscribers' *sessions* that took place from 1st July to 31st October, 2013. It is important to enhance the concept of session in our work. In the 3G standards 3GPP or 3GPP2, a session is created when the radio channel is allocated to a subscriber, as soon as he has data to be sent. Radio channel might be seen generically as a radio resource, e.g., time slot, code, or frequency. The session is finished by the network after a period of dormancy presented by the subscriber, which is configurable and typically set from 5 to 30 seconds [15]. The studied dataset

contains more than 1 billion sessions and each of them has the following information fields : (1) amount of upload and download volumes (in KiloBytes) during the session ; (2) session duration in seconds ; and (3) timestamp indicating when the session starts.

Due to the routinary behaviour of people [2] and the large scale of the dataset, it suffices to study a subset of the whole dataset in order to capture the daily behaviour of subscribers. Indeed, our analysis shows that there is low variability on subscribers' activity among the same hours on different days. Therefore, we have selected one week to more deeply assess the subscribers' behaviour. The studied week spans from 25th August to 31st August 2013 and contains information of about 2.8 million smartphone devices (the highest number of devices among the dataset weeks) and activity that totalizes 104 million sessions. This week has no special days or holidays and it is out of the Mexican preferred vacation period, which spans from early July to mid-August. From the data contained in this week, we have seen an enormous frequency of outliers on the first hour of all days, likely generated by the probe when the data collection was done. Therefore, we have discarded the data from midnight to 1am of all days in the following analysis. This does not affect our methodology that is indifferent to the amount of valid hours that the dataset provides.

Selecting a subset of one week allows us to better assess the subscribers' behaviour but it is important to emphasize that we will use the whole dataset later to evaluate our mobile traffic generator. Moreover, contrarily to datasets only describing CDRs, the richness of the considered dataset allows us to study and model detailed and realistic data traffic demands over time.

In the following, we start our analysis by studying the behavior of mobile subscribers in terms of traffic they generate and their activity on the temporal scale.

3.1 Traffic dynamics

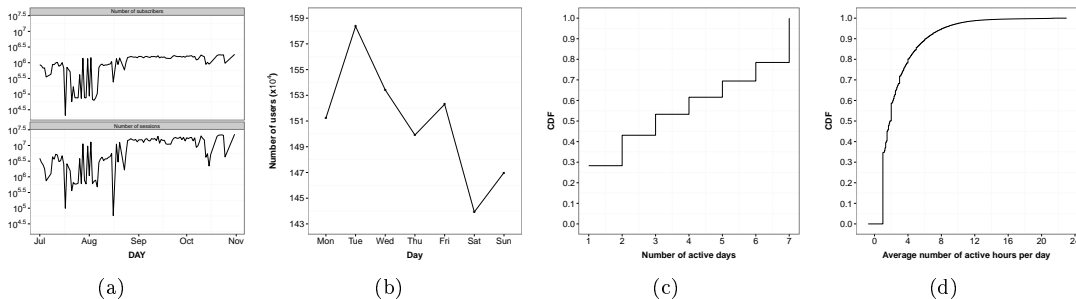


FIGURE 1 – (a) Number of subscribers and sessions on the whole dataset. (b) Number of subscribers per day generating traffic. (c) CDF of number of days in which subscribers generate traffic. (d) CDF of number of hours in which subscribers generate traffic per day during the week.

Fig. 1(a) shows the total number of subscribers and the total number of sessions from the whole dataset. As expected, *the number of subscribers and number of sessions are highly correlated*. It is possible to see a similarity on the shape of the curves for both parameters. Indeed, Spearman's correlation between number of users and number of sessions is 98%.

In Fig. 1(b), we present the number of subscribers that generated traffic on each of the days throughout the selected week (recall that the selected week is 25th August to 31st August 2013). *The day-wise variation of the number of active subscribers is essentially decreasing as the week progresses*. The difference between the weekdays and the weekend in terms of active subscribers is considerable; the highest difference is 10% which is obtained by comparing Tuesday with Saturday. As expected, *on average, the number of active subscribers are higher during the*

weekdays than during the weekend (also observed in [16]). In the studied week, this average difference is 5%.

Fig. 1(c) shows the CDF of the number of active days of the subscribers within the week (a subscriber is said to be active on some day if she generates some traffic on that day). *It is interesting to see that 22% of the subscribers generated traffic on all days, while 29% of the subscribers generated traffic only on one day of the week.* Also, 53% of the subscribers generated traffic on three or less days during the week. Similar percentages were measured from a different dataset and reported in [16].

Similarly, in Fig. 1(d) we show the CDF of the average number of active hours of the subscribers per day. We see that *most of subscribers generate traffic on few hours during the day.* Indeed, on an average 80% of the subscribers generate traffic for up to 4 hours each day. If we consider a longer period, e.g., for up to 6 hours, the number of such subscribers reaches 90%.

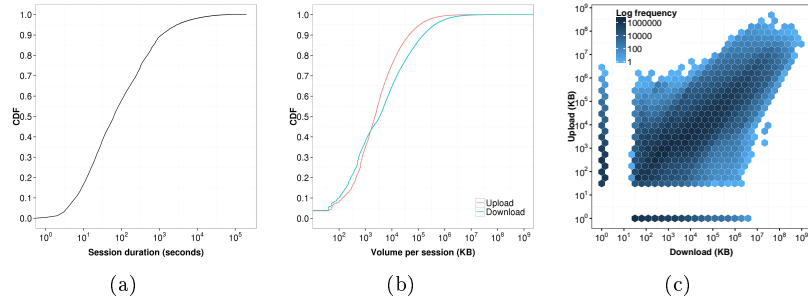


FIGURE 2 – (a) CDF of session duration in seconds per subscriber during the week. (b) CDF and (c) bin plot of the upload and download volume during the week.

Fig. 2(a) presents the CDF of session duration per subscriber during the week. *We see a median usage of 63 seconds of session and a significant variation in the amount of sessions used by different subscribers.* Interestingly, most of the sessions present short duration and few subscribers (less than 1%) use more than 6 hours of session during the week. In particular, the duration of 58% of the sessions is at most 100 seconds, while 90% of the sessions lasts for up to 15 minutes (similar behavior was reported in [16]).

Fig. 2(b) shows the CDFs of the average upload and download volumes of traffic generated per session. Observe that both the upload and download CDFs are similar : e.g., 35% and 38% of the sessions, respectively, present upload and download volume of up to 1 MB. On the other hand, 6% and 13% of the sessions present more than 100 MB for uploaded and downloaded volume, respectively. *We can observe that the median traffic load generated by typical subscribers is not significant while there are a small number of "heavy hitters" that consume a significant amount of network resources.*

Fig. 2(c) shows the *hexagonal bin plot* [17] of uploaded and downloaded volumes per session during the week. The intensity of a bin represents the frequency of sessions that generated upload and download volumes that lie within the bin. *The hexagonal bin plot reveals an uphill pattern from left to right, indicating a positive linear relationship between the per-session uploaded and downloaded volumes.* That is, if the amount of downloaded traffic is higher in a section, we can expect the uploaded volume to be higher as well. Indeed, the Spearman's correlation coefficient between per-session upload and download traffic is 88%. Aside from the uphill behaviour, we also observe two groups of bins forming straight lines, one close to each of the axis. Bins close to the x-axis are due to sessions that present a small upload volume, e.g., around 1 KB, and significantly

higher amount of download. Those are likely sessions in which subscribers use streaming media sites, e.g., Youtube, that typically use Real Time Protocol (RTP). RTP does not require the subscribers' device to generate confirmation packets, which justifies the small amount of uploaded volume. On the other hand, bins close to the y-axis represent sessions with small amount of download and comparably higher amount of upload. That is probably due to upload of media formats, e.g., photos on Facebook or videos on Youtube.

Owing to the high correlation between the upload and download volumes, in our evaluation and traffic modelling, we take into consideration the total volume per session, i.e., the sum of the upload and download volumes during the session.

3.2 Temporal dynamics

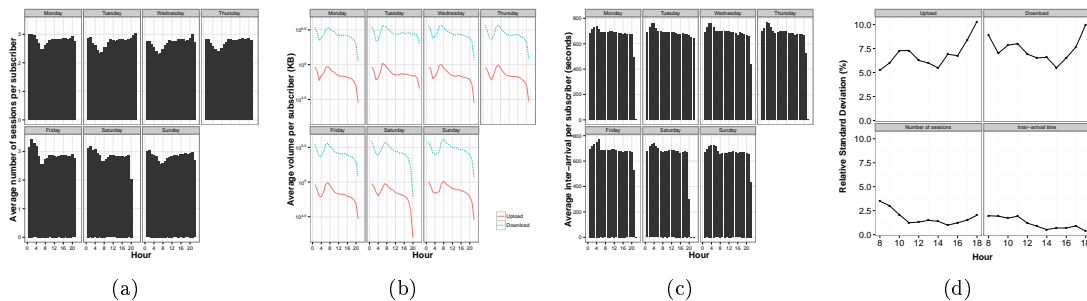


FIGURE 3 – (a) Average number of sessions per user during the week. (b) Volume of traffic for upload and download during the week. (c) Inter-arrival time per subscriber during the week. (d) Relative Standard Deviation per parameter.

It is common knowledge that some hours tend to be more active than others when it comes to our routinary daily activities. In this context, active (or peak) hours present high frequency of requests and volume of traffic, while non-active (or non-peak) hours present less traffic demands and volume. Indeed, Figs. 3(a), 3(b) and 3(c) show three parameters and their hourly dynamics during the week. Two features are important to highlight : *First, there is a repetitive behaviour during different days at the same hours. Second, there are peak and non-peak hours when it comes to subscribers' traffic demands.* In the following, we discuss these features and measure how repetitive their behaviour is. We further develop the idea of peak and non-peak hours for the users' activity in our traffic model.

Fig. 3(a) shows the average number of sessions per subscriber on each hour during the studied week. The result shows a clear gap on the average number of sessions from 4am to 8am. *On the end of late night and beginning of the day subscribers tend to perform less sessions.* This is consistent with diurnal human activity patterns. The number of sessions generated from 4am to 8am is 10% less when compared with that generated during the rest of the day. Furthermore, the total number of sessions from 9am to 3am is 47% higher than from 4am to 8am. Such behavior repeats over all days of the week.

Fig. 3(b) shows the upload and download session volumes per user during the week. *Similarly to the number of sessions (Fig. 3(a)), it is possible to see both the gap between 4am to 8am and the day-wise similarity.*

Fig. 3(c) shows the inter-arrival time (IAT), i.e., the difference between the arrival timestamps of subsequent sessions of the same subscriber. The high IAT shown from 4am to 8am is a complementary behaviour to the low average number of sessions on the same hours present in

Fig. 3(a). This is expected, and is due to the fact that *longer inter-arrival times results in less number of sessions on average*.

In summary, these last three results show *a high similarity on number of sessions, volume of traffic, and inter-arrival time traffic parameters, when compared day-wise. Indeed, all traffic parameters have similar per-hour values on different days, even comparing weekdays and weekends*. We measure the day-wise variability on subscribers' behaviour using the Relative Standard Deviation (RSD). RSD is the absolute coefficient of variation (CV), which is defined as the ratio of the standard deviation σ to the mean μ . Fig. 3(d) shows the per-parameter average RSD, which considers the hour-wise variation from all 7 days during Mexican working hours (i.e., from 8am to 6pm). It is possible to see that the maximum variability is small for all parameters : 3.4% for number of sessions, 1.9% for inter-arrival time, 10.3% and 9.9% for upload and download volumes, respectively. We have also calculated the maximum RSD of the parameters when compared day-wise, i.e., we have calculated the maximum RSD of each parameter on all hours of each day. The results show 4% for number of sessions, 2% for inter-arrival time, 16% and 15% for upload and download volumes, respectively. Therefore, we can conclude that, *on the studied dataset, the parameters from the same hours on different days present less variability than the parameters within the same day on different hours*.

The similarity of the temporal activity patterns among different days of the week is due to people's natural routinary behaviour. Therefore, *we select one day (namely, Wednesday) of the week to perform our extensive per-hour analysis and distinguish different profiles of users*.

4 Subscriber profiling methodology

This section describes the methodology of our work. We first define a limited number of profiles generated according to two traffic parameters : traffic demands (i.e., volume of traffic) and activity behaviour (i.e., number of sessions). Such parameters are extracted from a sample set of the considered dataset describing subscribers' traffic demands. Once such parameters are extracted from the sample set, a set of profiles is generated in order to describe the behaviour of subscribers. The profile definition procedure is performed through three phases. First, the similarity metric between all pairs of subscribers on the sample set is measured according to the two traffic parameters. Second, subscribers are clustered by their similarity into a limited number of clusters, also representing profiles. The third phase allows to classify the remaining additional subscribers of the dataset into the previous defined profiles. This profiling procedure results in typologies of subscribers based on their traffic dynamics. These different phases are detailed in the remainder of this section.

4.1 Similarity computation

Although we later evaluate our methodologies for a day within the week, our development in this section can hold in general for any time interval D chosen from the week. For a given time interval D , let \mathbb{S} be the set of all subscribers that generate some traffic during D , and $\mathbb{S}' \subseteq \mathbb{S}$ be a randomly selected sample of subscribers from \mathbb{S} . Our objective is to partition the subscribers in \mathbb{S}' into a set of *clusters* \mathbb{P} , such that subscribers belonging to the same cluster are "similar" in terms of traffic demands. We use Euclidean distance to measure the *similarity* between two subscribers [18]. We then *classify* the remaining users in \mathbb{S} (i.e., $\mathbb{S} - \mathbb{S}'$) into various clusters in \mathbb{P} . In this work, we develop similarity comparison according to *volume of traffic* and *number of sessions*. These traffic parameters allow us to make a comparison between two different subscribers; our clustering and classification algorithms (discussed in the next section) are also based on these parameters.

Each subscriber i can be effectively represented as a sequence of sessions generated by i . Let t_k^i denote the time instant at which the k -th session of subscriber i begins. Let v_k^i be the volume of traffic (both upload and download) generated by subscriber i during the k -th session. However, this very fine representation of a subscriber is costly in terms of the memory and processing time required. To overcome this drawback, we divide D into time slots of length T . Thus, there are $\frac{D}{T}$ number of time slots. The notion of time slots allow us to collect together all sessions occurring within t .

For subscriber $i \in \mathbb{S}'$, let τ_t^i denote the set of all sessions starting within time slot t , i.e., $\tau_t^i = \{k : (t-1)T \leq t_k^i \leq tT\}$. Now, the volume of traffic generated by subscriber i , in time slot t , is given by

$$V_t^i = \sum_{k \in \tau_t^i} v_k^i. \quad (1)$$

Similarly, the number of sessions generated by subscriber i in time slot t can be written as

$$N_t^i = \sum_k \mathbb{I}(k \in \tau_t^i) \quad (2)$$

where $\mathbb{I}(k \in \tau_t^i) = 1$ if $k \in \tau_t^i$; 0 otherwise. Thus, to obtain N_t^i we simply count the sessions of subscriber i that begin inside time slot t .

Using the above expressions, it is now easy to obtain the total volume and the total number of sessions generated by subscriber i during D : $\vartheta^i = \sum_{t \in D} V_t^i$ and $\eta^i = \sum_{t \in D} N_t^i$. Finally, we define the *traffic volume similarity* between two subscribers i and j as the difference between the total volumes generated by these users, i.e.,

$$w_{ij}^\vartheta = \|\vartheta^i - \vartheta^j\|. \quad (3)$$

The *number of sessions similarity* can be similarly defined as :

$$w_{ij}^\eta = \|\eta^i - \eta^j\|. \quad (4)$$

Using the subscribers in \mathbb{S}' as the vertices, and using either $w_{i,j}^\vartheta$ or $w_{i,j}^\eta$ as the edge weights, we obtain a complete graph $G(\mathbb{S}', \mathbb{E})$, which is given as input to our clustering algorithm to obtain different clusters in \mathbb{P} . Then, after classifying the remaining users (i.e., $\mathbb{S} - \mathbb{S}'$), we use V_t^i and N_t^i to further classify the time slots into peak and non-peak.

4.2 Subscriber clustering and classification

Instead of a-priori fixing a value for the number of profiles (i.e., clusters) $|\mathbb{P}|$, our goal is to obtain from the data, how many profiles are needed to best represent the subscribers' traffic activities. For this purpose, we use an hierarchical clustering algorithm that iteratively aggregates vertices from the similarity graph $G(\mathbb{S}', \mathbb{E})$ into larger clusters, according to a dendrogram structure [19]. The hierarchical clustering algorithm we choose is the *Average Linkage clustering method*, also known as *Unweighted Pair Group Method with Arithmetic Mean (UPGMA)* [19]. UPGMA starts by first considering each vertex (subscribers in \mathbb{S}' in our case) of the given graph as a cluster (i.e., singleton clusters). At each iteration, it computes the distance (using the edge weights between vertices given by Eq. (3) or Eq. (4)) between all pairs of clusters, and then merges the closest two clusters. Thus, in our context, it merges together the two clusters that are more similar in terms of the traffic demands. If the algorithm is not stopped, it finally simply yields a single cluster containing all the vertices. Thus, it is important to find where UPGMA

should stop its merging process, yielding the best number of clusters, i.e., *the best separation among the groups of usage pattern from subscribers*. To that end, we use several *stopping rules* (or stopping criteria). A stopping rule, during each iteration of the hierarchical clustering algorithm (or each level of the dendrogram), gives a measure of how well separated the clusters are, based on which one can decide the best number of clusters to use.

In the literature, there are several stopping rules [20]. Contrarily to related works that have implemented and applied very few of them [16] and in order to avoid to be biased by a specific criteria, we have implemented and used 23 stopping rules, namely Ball-Hall index [20], Beale index [20], Cubic Clustering Criterion index [20], Calinski-Harabasz index [20], C-index [20], DB index [20], Duda index [20], Dunn index [21], Frey index [20], Friedman index [20], Hartigan index [20], Krzanowski-Lai index [22], Marriot index [20], McClain-Rao index [20], Pseudot2 index [20], Ratkowsky-Lance index [20], Rubin index [20], Scott-Symons index [20], SDbw index [18], SD index [23], Silhouette index [24], TraceW index [20], TraceCovW index [20].

For the sake of illustration, we will briefly describe the C-Index [20] stopping rule here. C-Index is defined as $C = (S - S_{min}) / (S_{max} - S_{min})$, where : (1) S is the sum of all distances between pairs of users in the same cluster over all clusters, (2) S_{min} and S_{max} are the sum of the smallest and the largest distances respectively, for all pairs of users, over all clusters. In our context, it compares the distances among the considered traffic parameters. According the C-Index, the lower the value of the index, the better the clustering. In this way, the number of profiles producing the lowest C-Index index value is the one that grants the best separation among clusters.

Fig. 4(a) shows the C-Index and the Beale index values as a function of the number of clusters, when traffic volume similarity is used to compute the distance between pairs of users. while C-Index considers choosing the best number of clusters based on its minimum index value, Beale uses the maximum index value. Thus, in both index, the best number of clusters is 2 according to Fig. 4(a).

Similarly, each other 21 implemented stopping rules listed above define their best number of clusters to be used. In Fig. 4(b), we present the frequency of the best number of clusters, while profiling subscribers using traffic volume similarity. It condensates in a histogram the result of the 23 stopping rules. It shows that 8 stopping rules recommend 2 as the best profiles when clustering subscribers by their traffic volumes.

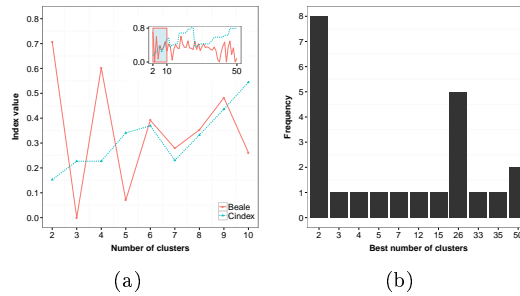


FIGURE 4 – (a) Beale index and C-Index values and respective number of clusters, when using the traffic volume similarity (b) Histogram of best number of clusters indicated by the assessed stopping rules.

Profiling occurs then in four stages : (1) building a similarity graph with S' subscribers, (2) hierarchically clustering it using a similarity metric, (3) determining the best number of clusters $|\mathbb{P}|$, i.e., profiles relying on the stopping rules, and (4) classifying $S - S'$ remaining unclassified

subscribers in the previous defined clusters. In the fourth stage, we use the *k-means algorithm* as the classification technique. It is worth mentioning, we calculate the clusters centroids (means) obtained from the hierarchical clusters and use them on the first iteration of the k-means algorithm. This is an important information because the centroids obtained from the hierarchical clustering algorithm are likely to be better positioned than the k-means originally bootstrapped initial centroids, which are based on randomly selected positions.

These four stages are performed in two rounds. In the first round, the graph $G(\mathbb{S}', \mathbb{E})$ weighted according to the traffic volume similarity (Eq. (3)) is used at the hierarchical clustering. The best number of "traffic volume"-based clusters is then determined : According to results shown in Fig. 4(b), $|\mathbb{P}| = 2$ weighted subgraphs $\{G_1(\mathbb{S}'_1, \mathbb{E}), G_2(\mathbb{S}'_2, \mathbb{E})\}$ are created. At the end of the first round, the final classification takes place. The next execution round initiates with the a new hierarchical clustering being performed inside each initially defined "traffic volume"-based cluster. This time G_1 and G_2 are weighted according to the number of sessions similarity (Eq. (4)). Finally, for each of these two initial clusters, two "number of sessions"-based clusters are defined after the second round of stopping rules execution, totalizing four subscribers profiles. Due to space constraints, we will not show these last stopping rules results. The second round ends with the classification of the remaining subscribers into the four defined profiles. Next section better details our subscriber profiling.

4.3 Subscriber profiles

To obtain the profiles for our dataset, we set D as 27th of August, which contains information of about 1.5 million smartphone devices, and randomly sampled 10000 subscribers (thus, $|\mathbb{S}'| = 10000$). D is a normal day with no special event or holiday and we divide it into time slots of duration T . Time slots help to understand the general behaviour of a certain period of time in D . Higher the number of time slots, shorter is their duration and vice-versa. Very short time slots, e.g., 1 minute, may lead to an analysis with fewer sessions per time slot, hindering the identification of subscribers' behaviour per slot. Very large time slots, e.g., 12 hours, may lead to a general view of the sessions, so that it is difficult to obtain a good quality assessment of the traffic dynamics. Thus, for our evaluation we choose a "moderate" value of 1 hour as the time slot duration. However, the optimal size of the time slot is still an open problem [25].

Our profiling methodology resulted in *four profiles*, and we have named them as follows : Light Occasional (LO), Light Frequent (LF), Heavy Occasional (HO) and Heavy Frequent (HF). *Light* profiles contain subscribers that generate up to 20 GB of data during the day, while *Heavy* profiles have subscribers that generate more than 20 GB of traffic during the day. Likewise, *Occasional* profiles contain subscribers that are generating up to 278 connection sessions, whereas *Frequent* profiles contain users generating more than 278 connections per day. Table 1 shows the characteristics of each of the profiles.

TABLE 1 – Characteristics of the resulting profiles

	Light		Heavy	
	Occasional	Frequent	Occasional	Frequent
Volume	29 KB to 20 GB		21 GB to 625 GB	
N° of subscribers	1489242		27659	
N° of sessions	1 to 278	279 to 8737	1 to 278	279 to 8737
N° of subscribers	1486496	2746	27593	66

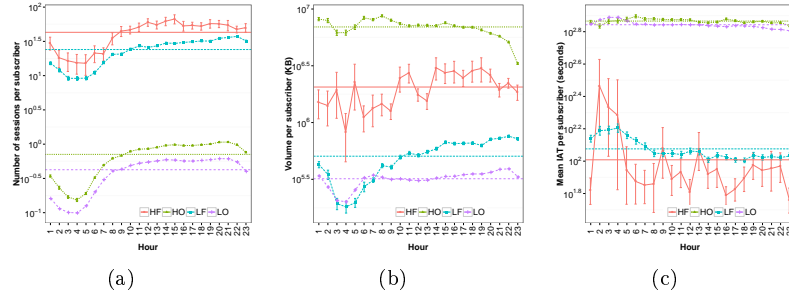


FIGURE 5 – (a) Mean inter-arrival per class. (b) Number of sessions per class. (c) Volume of traffic per class.

In Fig. 5, we show the dynamics of the traffic parameters per subscribers' class per hour. Fig. 5(a), 5(b), and 5(c) corresponds to the number of sessions, volume of traffic, and the mean inter-arrival time, respectively; the error bars correspond to a 95% confidence interval. For each time slot, the volume of traffic and number of sessions are calculated using Eq. (1) and Eq. (2), respectively. For each subscriber i , the average inter-arrival time in time slot t is obtained using the following expression :

$$IAT_t^i = \frac{\sum_{k \in \tau_t^i} (t_{k+1}^i - t_k^i)}{N_t^i}, \quad (5)$$

where τ_t^i is the set of all sessions of subscriber k that lie with the time slot t . Similar to ϑ^i and η^i , we define the average inter-arrival time for the entire D as $\zeta^i = \sum_{t \in D} IAT_t^i$.

From Fig. 5, we can see that our methodology well separates the profiles, i.e., the *occasional* and *frequent* subscribers have their values clearly separated. For each curve in these plots, we have also shown a *horizontal line that represents the respective mean value* (where the mean is taken over all time slots). Given the mean values, we classify, for each profile of subscribers and for each parameter (number of sessions, traffic volume, and IAT), the hours above the mean as *peak hours*, and hours below the mean as *non-peak hours*.

5 Measurement-driven traffic modeling

The goal of the traffic model is to generate synthetic subscribers, whose usage pattern is consistent with the observations made about the real subscribers in the previous section. Recall that subscribers belonging to different profiles (HO, HF, LO, and LF) have their own specificities in terms of *when* the sessions are generated during the day, and the *amount of* volume generated during each session. Furthermore, each profile of subscribers have different behaviour during *peak and non-peak hours*. Thus, to obtain a fine grained model it is important to take into account all the above considerations, while simulating a synthetic trace. In the following, we describe how we merge all the above considerations to obtain a measurement-driven mobile data traffic modeling.

5.1 Fitting empirical distributions

Using the original subscribers' data, we first study for each profile in peak and non-peak hour, the empirical distribution functions (i.e., CDF) of the traffic parameters (see Figs. 6 and 7) :

the number of sessions generated, the traffic volume associated with each of these sessions, and the inter-arrival times between the sessions. For instance, the empirical distribution function of "total volume for HF users in peak hours" is obtained from the set of all V_t^i (Eq. (1)) such that $i \in \mathbb{S}$ is an HF subscriber and t is peak hour. The empirical distribution functions of the number of sessions and the inter-arrival time for any combination of profile and hour-type (peak or non-peak), can be similarly generated using N_t^i (Eq. (2)) and IAT_t^i (Eq. (5)), respectively.

Once the CDFs are obtained, using statistical tests we estimate the set of distributions that best fit them. From this set, we then select the closest distribution function to the respective CDF to be *the function to use at the traffic usage pattern generation for the corresponding profile and type of hour*. More specifically, when considering the volume of traffic and the inter-arrival time parameters (i.e., consisting of continuous values) of a certain profile and hour, the Kolmogorov-Smirnov statistic test [26] is used. The test estimates the parameters for a set of continuous distributions (namely, Log-normal, Gamma, Weibull, Logis, and Exponential) that best fit the corresponding empirical distribution function. Similarly, when considering the number of sessions parameter (i.e., consisting of discrete values) of a certain profile and hour, the Chi-squared statistic test [27] is used to estimate the best fitting parameters for a set of discrete distributions (Negative binomial, Geometric, and Poisson). In both cases, after getting the sets resulted from the fitting tests, we select the distribution functions that best fit each corresponding CDF.

Tables 2, 3, and 4 list the best fitted distribution functions along with their parameters for all possible combination of profile and hour-type pair, for respectively, number of sessions, traffic volume and inter-arrival time parameters. For Negative-binomial distribution, n is the size parameter and p is the probability parameter. For Gamma distribution, α indicates the shape parameter and β is the rate parameter. For Weibull distribution, k is the shape parameter and λ refers to the scale parameter. For Log-normal distribution, σ represents the shape parameter and μ is the scale parameter. For Gamma, Weibull and Log-normal, x_0 is the location parameter.

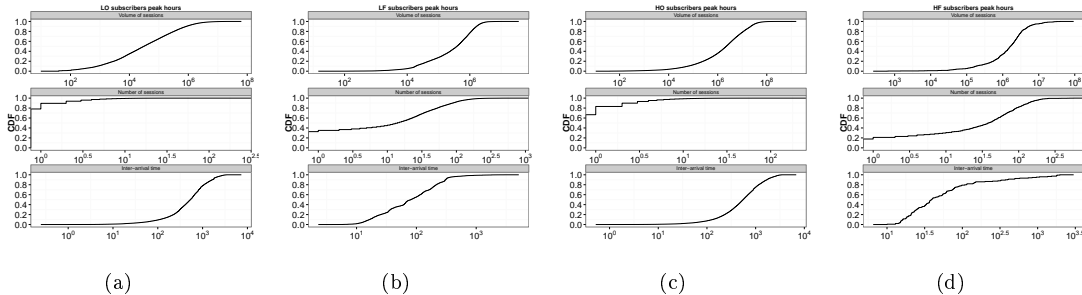


FIGURE 6 – (a) Traffic parameters for LO users in peak hours. (b) Traffic parameters for LF users in peak hours. (c) Traffic parameters for HO users in peak hours. (d) Traffic parameters for HF users in peak hours.

5.2 Synthetic subscriber generation

Generating a synthetic subscriber will first require us to generate a profile type (HO, HF, LO or LF) for the subscriber. Profile types are assigned randomly, based on the distribution of profiles population observed in the real data. For instance, from Table 1 we see that 97.995% of the subscribers belong to LO profile, and thus with probability $q_{LO} = 0.97995$ we assign LO profile to a synthetic user. Similarly, the probabilities of other profiles are : $q_{LF} = 0.00181$,

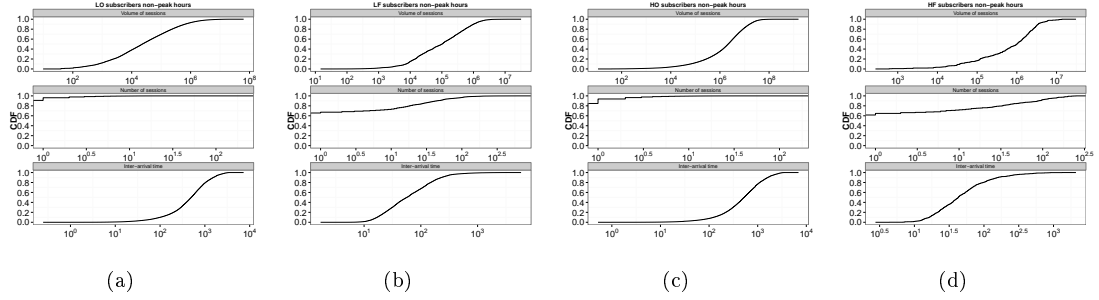


FIGURE 7 – (a) Traffic parameters for LO users in non-peak hours. (b) Traffic parameters for LF users in non-peak hours. (c) Traffic parameters for HO users in non-peak hours. (d) Traffic parameters for HF users in non-peak hours.

$q_{HO} = 0.01819$, and $q_{HF} = 0.00004$. We will refer to $q = (q_{LO}, q_{LF}, q_{HO}, q_{HF})$ as the *profile pmf*, or probability mass function.

We now briefly describe our procedure for generating a synthetic subscriber presented in Algorithm 1. *We first randomly generate a profile type for a subscriber i using the profile pmf q . After obtaining the profile type, for a given hour t , we randomly sample values for each traffic parameter according to the corresponding fitted distribution functions.*

In more detail, the algorithm requires one parameter which is the number of synthetic users to be generated (line 1). The result of the generation is a list of sessions per user (line 25). Each synthetic user session contains two fields: (1) volume of traffic and (2) arrival timestamp. For each subscriber i and time slot t , we sample a number of sessions N_t^i , the mean inter-arrival time IAT_t^i , and the average session volume V_t^i from the appropriate distributions (i.e., the fitted distribution corresponding to the profile and hour-type pair) listed in Tables 2, 4, and 3, respectively (lines 5 to 7). The volume per session v_k^i (for $k \in \tau_t^i$, see Section 5) is then equal to the sampled value V_t^i divided by the sampled number of sessions N_t^i . The initial timestamp of each session in hour t is then computed according to the sampled inter-arrival time IAT_t^i and number of session N_t^i for that hour (lines 11 to 18). By varying t over the 24 hours in a day, we obtain a synthetic subscriber traffic for one day.

TABLE 2 – Number of sessions : distributions and parameters

Number of sessions			
Hour	Profile	Distribution	Parameters
Peak	HO	Neg-binomial	$n = 0.1252, p = 0.11$
	HF		$n = 0.7831, p = 0.01$
	LO		$n = 0.0915, p = 0.14$
	LF		$n = 0.4367, p = 0.01$
Non-Peak	HO	Neg-binomial	$n = 0.0477, p = 0.11$
	HF		$n = 0.2282, p = 0.01$
	LO		$n = 0.0317, p = 0.12$
	LF		$n = 0.1571, p = 0.01$

TABLE 3 – Session volume : distributions and parameters

Session volume			
<i>Hour</i>	<i>Profile</i>	<i>Distribution</i>	<i>Parameters</i>
Peak	HO	Weibull	$k = 0.54, \lambda = 4646633.6, x_0 = 39$
	HF	Weibull	$k = 0.84, \lambda = 2540145.8, x_0 = 613$
	LO	Gamma	$\alpha = 0.10, \beta = 0.0000002, x_0 = 29$
	LF	Gamma	$\alpha = 0.67, \beta = 0.000001, x_0 = 40$
Non-Peak	HO	Weibull	$k = 0.58, \lambda = 4096489.1, x_0 = 40$
	HF	Gamma	$\alpha = 0.61, \beta = 0.0000003, x_0 = 613$
	LO	Gamma	$\alpha = 0.08, \beta = 0.0000002, x_0 = 29$
	LF	Gamma	$\alpha = 0.28, \beta = 0.0000007, x_0 = 40$

TABLE 4 – Session mean inter-arrival times : distributions and parameters

Session mean inter-arrival time			
<i>Hour</i>	<i>Profile</i>	<i>Distribution</i>	<i>Parameters</i>
Peak	HO	Gamma	$\alpha = 1.3060, \beta = 0.001, x_0 = 1$
	HF	Log-normal	$\sigma = 4.0106, \mu = 1.2114, x_0 = 10$
	LO	Gamma	$\alpha = 1.2799, \beta = 0.001, x_0 = 0.5$
	LF	Weibull	$k = 0.9173, \lambda = 135, x_0 = 3.9$
Non-Peak	HO	Gamma	$\alpha = 1.2679, \beta = 0.001, x_0 = 1$
	HF	Log-normal	$\sigma = 3.8552, \mu = 0.9196, x_0 = 4.3$
	LO	Gamma	$\alpha = 1.2799, \beta = 0.001, x_0 = 0.5$
	LF	Log-normal	$\sigma = 4.1174, \mu = 1.0291, x_0 = 3$

5.3 Synthetic traffic model evaluation

In order to evaluate our traffic modeling, we generate a synthetic dataset and compare it with the original dataset. Towards this goal, we first generate a set \mathbb{R} of synthetic subscribers, where $|\mathbb{R}| = |\mathbb{S}|$, for one day of traffic. The synthetic dataset contains for each session of a subscriber i and at hour t : (1) the volume in KiloBytes generated and (2) the initial timestamp of the session. We assess how consistent the synthetic traffic is by comparing the distributions of the various parameters between the original and the synthetic datasets.

For this, we use the Bhattacharyya (BH) measure [28]. It quantifies the similarity between two discrete or continuous probability distributions. Let $p(i)$ and $p'(i)$ be two pmfs, i.e., $\sum_{i=1}^N p(i) = \sum_{i=1}^N p'(i) = 1$. Then, BH measure is formally defined as $\rho(p, p') = \sum_{i=1}^N \sqrt{p(i)p'(i)}$. However, BH measure is not a distance metric since it does not satisfy all the metric axioms. Therefore, [29] proposes an alternative distance metric based on the BH distance which is formally defined as $d(p, p') = \sum_{i=1}^N \sqrt{1 - p(i)p'(i)}$. Note that, $d(p, p')$ exists for all discrete distributions and it is equal to zero iff $p = p'$. We use d in order to measure the similarity between the original dataset and the synthetic dataset.

Let \mathbb{D} denote a set of different time periods including D and the synthetic day denoted as D' . \mathbb{D} also contains each day from 1st July to 31st October, i.e., the whole dataset. Let p_v^e denote the PDF (Probability Distribution Function) of the total volume generated by a subscriber active in day e , formally defined as $p_v^e(x) = \sum_{i \in e} \mathbb{I}(v^i = x) / |\{i \in e\}|$. For instance, in Fig. 8(a) we have

depicted the CDFs corresponding to the PDFs p_{θ}^D and $p_{\theta}^{D'}$. We can observe an *almost complete overlap of the two CDFs due to high similarity between the real trace and the synthetic trace*.

To evaluate our traffic model, we first compute $d(p_{\theta}^D, p_{\theta}^{D'})$, the distance between total volume distribution of the original day and the synthetic day. Then, we compute $d(p_{\theta}^D, p_{\theta}^e)$, $e \in \mathbb{D}$ but $e \neq D$, the distance between the original day and remaining days in the original trace. We obtain similar distances for p_{η}^e and p_{ζ}^e for $e \in \mathbb{D}$, which are respectively, the PDFs of the total number of sessions and average inter-arrival time by a subscriber active in day e . Finally, for each distribution, we have also computed the mean and the confidence interval (95%) of the distances between the original day and the remaining days. In Fig. 8(b), we show the $d(p_{\theta}^D, p_{\theta}^e)$ distances. Also shown in Fig. 8(b) (horizontal dashed line) is the $d(p_{\theta}^D, p_{\theta}^{D'})$ distance. Our first traffic model evaluation consists then in verifying if the $d(p_{\theta}^D, p_{\theta}^{D'})$ is within the confidence interval of the $d(p_{\theta}^D, p_{\theta}^e)$. As can be seen in Fig. 8(b), *for each distribution, the distance of the synthetic day (from the original) is within this confidence interval*.

Finally, we applied the profiling methodology described in Section 4 on the synthetic subscribers. By doing so, we classify them and compare the per-class traffic behaviour with the one created from the original dataset. Fig. 8(c) depicts the per-class behaviour for the volume of traffic per session for the classified synthetic subscribers. It is possible to see that this *result is coherent with the one for the original dataset* presented in Fig. 5(b). For instance, the behaviour for peak and non-peak hours is well defined and similar to the one from the original trace.

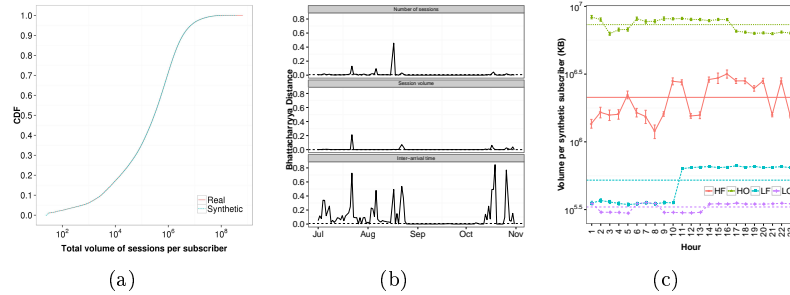


FIGURE 8 – (a) CDF of the total volume per session for real and synthetic subscribers (b) Per-parameter Bhattacharyya distances between original and synthetic trace (dashed line) in D , and between original trace in D and other days from the original trace (line) (c) Volume of traffic per class for synthetic subscribers.

6 Discussion and Work implication

An important aspect on network planning and management is to know what is the load it will be subjected to. Subscribers with different profiles impose, on certain cases, totally different demands to the network, and they may or may not be geographically far from each other. For instance, our dataset shows that the 276 thousand lightest subscribers generate similar amount of traffic as generated by the heaviest subscriber in the entire day.

A key importance of this work is to provide distributions to model workload characteristics of mobile subscribers' traffic demands and a framework on how to create a traffic generator out of it. Therefore, this work has implication in some areas related to network planning such as hotspot deployment [2].

Hotspot deployment generally denotes a problem in which the objective is to provide the best placement for hotspots respecting certain constraints. For instance, one may desire to deploy

a fixed amount of hotspots to maximize the amount of data offloaded from the network. The literature frequently presents the evaluation of hotspot deployment based on mobility datasets describing subscribers' trajectories. Although literature provides some mobility datasets, to the best of our knowledge none of them provides information of both mobility and traffic demands. Our traffic generator could be attached to the mobility datasets and it would allow to better exploit them.

7 Conclusions and next steps

In this paper we have first presented a characterization of a 4-month dataset that contains more than 1.05 billion session connections from about 6.8 million smartphone users. Moreover, we propose a framework that automatically classifies those users by their traffic demands into a limited number of profiles. Our approach takes advantage of repetitive user behaviour due to their daily routines. Furthermore, we have provided distributions that describe their traffic demands into peak and non-peak hours. Finally, from these distributions we create a traffic generator and evaluate the synthetic trace it generates. Our results show that the synthetic trace presents a consistent behaviour when compared to original dataset.

As future work, we aim to investigate models to describe sessions' transfer rate and duration. Additionally, we intend to apply and evaluate our traffic generator on different problems such as network planning.

Table des matières

1	Introduction	3
2	Related Work	4
3	Dataset	4
3.1	Traffic dynamics	5
3.2	Temporal dynamics	7
4	Subscriber profiling methodology	8
4.1	Similarity computation	8
4.2	Subscriber clustering and classification	9
4.3	Subscriber profiles	11
5	Measurement-driven traffic modeling	12
5.1	Fitting empirical distributions	12
5.2	Synthetic subscriber generation	13
5.3	Synthetic traffic model evaluation	15
6	Discussion and Work implication	16
7	Conclusions and next steps	17

Références

- [1] Cisco. (2013, Feb.) Cisco visual networking index : Global mobile data traffic forecast update, 2013–2018.
- [2] E. M. R. Oliveira and A. C. Viana, “From routine to network deployment for data offloading in metropolitan areas,” in *Proc. of IEEE SECON*, Jun. 2014.
- [3] J. Candia, M. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabasi, “Uncovering individual and collective human dynamics from mobile phone records,” *Journal of Physics A : Mathematical and Theoretical*, vol. 41, 2008.
- [4] J. Wortham, “Cellphones now used more for data than for calls,” *New York Times*, May 2010.
- [5] R. Becker, R. Caceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, “A tale of one city : Using cellular network data for urban planning,” *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 18–26, Apr. 2011.
- [6] D. Naboulsi, R. Stanica, and M. Fiore, “Classifying call profiles in large-scale mobile traffic datasets,” in *Proc. of IEEE Infocom*, Apr. 2014.
- [7] A. Pawling, N. V. Chawla, and G. Madey, “Anomaly detection in a mobile communication network,” *Computational and Mathematical Organization Theory*, vol. 13, no. 4, pp. 407–422, 2007.
- [8] 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 *Proc. of the 1st Workshop on Urban Networking (ACM UrbaNe)*, Dec. 2012.
- [9] C. W. O. O. A. Abidogun, “A self organizing maps model for outlier detection in call data from mobile telecommunication networks,” in *Proc. of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, Aug. 2004.
- [10] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, “Mobile data offloading : How much can wifi deliver?” *Networking, IEEE/ACM Transactions on*, vol. 21, no. 2, pp. 536–550, April 2013.
- [11] A. Vaccari, L. Liu, A. Biderman, C. Ratti, F. Pereira, J. Oliveirinha, and A. Gerber, “A holistic framework for the study of urban traces and the profiling of urban processes and dynamics,” in *Proc. of Int. IEEE Conf. on Intelligent Transportation Systems (ITSC)*, Oct. 2009.
- [12] P. Paraskevopoulos, T. C. Dinh, Z. Dashdorj, T. Palpanas, and L. Serafini, “Identification and characterization of human behavior patterns from mobile phone data,” in *Proc. of NetMob*, May 2013.
- [13] R. M. Pulselli, P. Romano, C. Ratti, and E. Tiezzi, “Computing urban mobile landscapes through monitoring population density based on cellphone chatting,” *Int. Journal of Design and Nature and Ecodynamics*, vol. 3, 2008.
- [14] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, “Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate,” in *Proc. of Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.
- [15] Alcatel-Lucent, “Alcatel-lucent 9900 wireless network guardian,” White Paper, Dec. 2012.
- [16] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, “Understanding traffic dynamics in cellular data networks,” in *Proc. of IEEE Infocom*, Apr. 2011.
- [17] D. B. Carr, A. R. Olsen, and D. White, “Hexagon mosaic maps for displaying univariate and bivariate geographical data,” *Cartography & Geographical Information Systems*, vol. 19, pp. 228–236, 1992.

-
- [18] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, Dec. 2001.
- [19] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Scientific Bulletin*, vol. 28, pp. 1409–1438, 1958.
- [20] G. Milligan and M. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.
- [21] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.
- [22] W. J. Krzanowski and Y. T. Lai, "A criterion for determining the number of groups in a data set using sum-of-squares clustering," *Biometrics*, vol. 44, no. 1, pp. 22–34, Mar. 1988.
- [23] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, "Quality scheme assessment in the clustering process," in *Proc. of the 4th European Conf. on Principles of Data Mining and Knowledge Discovery*, Sep. 2000.
- [24] P. Rousseeuw, "Silhouettes : A graphical aid to the interpretation and validation of cluster analysis," *Elsevier Journal of Computational Applied Mathematics*, vol. 20, no. 1, pp. 53–65, Nov. 1987.
- [25] T. Hossmann, T. Spyropoulos, and F. Legendre, "Know thy neighbor : Towards optimal mapping of contacts to social graphs for DTN routing," in *Proc. of IEEE INFOCOM*, Mar. 2010.
- [26] R. B. D'Agostino and M. A. Stephens, *Goodness-of-Fit-Techniques*. CRC Press, Jun. 1986, vol. 68.
- [27] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine Series 5*, vol. 50, no. 302, pp. 157–175, 1900.
- [28] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.
- [29] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.

Algorithm 1 Traffic model algorithm

```

1: procedure GENERATE( $n$ )
2:   for each integer  $i$  from 1 to  $n$  do
3:      $class_i = \text{CLASS}()$ 
4:     for  $t$  in  $D$  do
5:       for  $Metric$  in  $\{N, IAT, V\}$  do
6:          $pMetric_i^t = \text{DISTR}(Metric, t, class_i)$ 
7:       end for
8:        $nrsess = \text{GET}(pN_i^t, 1)$ 
9:        $Arrivals_i = [ ]$ 
10:       $AvgIAT = \text{GET}(pIAT_i^t, nrsess)$ 
11:      for  $intarr$  in  $AvgIAT$  do
12:        if  $intarr == AvgIAT[0]$  then
13:           $arr = intarr$ 
14:        else
15:           $arr = AvgIAT[intarr - 1] + intarr$ 
16:        end if
17:         $\text{APPEND}(Arrivals_i, arr)$ 
18:      end for
19:       $Volumes_i = \text{GET}(pIAT_i^t, nrsess)$ 
20:    end for
21:     $Sessions_i = [ ]$ 
22:    for  $arrival, volume$  in  $Arrivals_i, Volumes_i$  do
23:       $session = \text{SESSION}(arrival, volume)$ 
24:       $\text{APPEND}(Sessions_i, session)$ 
25:    end for
26:    return  $Sessions_i$ 
27: end for
28: end procedure

```



**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 0249-6399