Mobile Data Traffic Modeling: Revealing Temporal Facets

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

Mobile Data Traffic Modeling: Revealing Temporal Facets

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

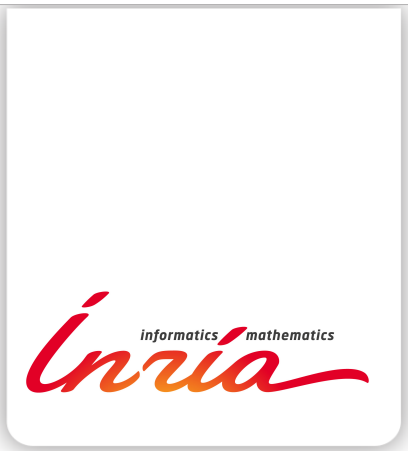
RESEARCH REPORT

N° 8613

October 2014

Project-Team Infine

ISSN 0249-6399 ISRN INRIA/RR--8613--FR+ENG



Mobile Data Traffic Modeling: Revealing Temporal Facets

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

Équipe-Projet Infine

Rapport de recherche n° 8613 — version 5 — version initiale October 2014 — version révisée Juin 2015 — 32 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 modèle 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

Mobile Data Traffic Modeling: Revealing Temporal Facets

Abstract: Using a large-scale dataset collected from a major 3G network in a dense metropolitan area, this paper presents the first detailed measurement-driven model of mobile data traffic usage of smartphone subscribers. Our main contribution is a synthetic, measurement-based, mobile data traffic generator capable of simulating traffic-related activity patterns for different categories of subscribers and time periods for a typical day in their lives. We first characterize individual subscribers routinary behaviour, followed by a detailed investigation of subscribers' temporal usage patterns (i.e., "when" and "how much" traffic is generated). We then classify the subscribers into six distinct profiles according to their usage patterns and model these profiles according to two daily time periods: peak and non-peak hours. We show that the synthetic trace generated by our data traffic model consistently replicates a subscriber's profiles for these two time periods when compared to the original dataset. Broadly, our observations bring important insights into network resource usage. We also discuss relevant issues in traffic demands and describe implications in network planning and privacy.

Key-words: networks, traffic model, routine

1 Introduction

Smartphone devices provide today the best means of gathering users information about content consumption behavior 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. The *understanding of users' mobile data traffic demands* is of fundamental importance when looking for solutions to manage the recent boost up of mobile data usage [1, 2, 3] and to improve the quality of communication service provided, favoring the proliferation of pervasive communication. Hence, the definition of a *usage pattern* can allow telecommunication operators to better forseen future demanded traffic and consequently, to better (1) deploy data offloading hotspots or (2) timely plan network resources allocation and then, set subscription plans.

Contrarily to most related work in the literature modeling call traffic (frequently referred as Call Detail Records (CDRs)), we characterize and model real mobile data traffic demands generated by smartphone subscribers. Although convenient and of frequent consideration, call records only provide an intuition of users activity in the network : voice calls and SMS. In addition, due to its sparsity in time [4], subscribers behavior in terms of call shows strong variations with time and day of the week [3] : a different behavior is found when data traffic is considered. Finally, call traffic does not describe the background traffic load automatically generated by current smartphone applications (e.g., email checks, synchronization). We thus claim that, since smartphones are now used more for data than for calls [5], the use of call records for investigating traffic demands is not enough for dimensioning network usages.

Our first contribution in this paper is *to profile urban mobile data traffic*. For this, we perform a *precise characterization of individual subscribers' traffic behavior clustered by their usage patterns*, instead of a network-wide data traffic view [6, 7, 8]. Note that the high dynamic behavior of individual subscribers (in terms of traffic demands and in time) and the use of large scale datasets make this task complex. In addition, for the purpose of quality of service testing of new applications, infrastructures, or network mechanisms, one needs a traffic generator that is capable of generating realistic synthetic traffic that "looks like" traffic found on an actual network. In this context, our second contribution is *to provide a way for synthetically, still consistently, reproducing usage patterns of mobile subscribers* – the first work in the literature to do so, to the best of our knowledge. The implications of this work are diverse, in particular, in resource allocation planning and testing, or hotspot deployment. When it comes to legal issues, it is also worth mentioning the unconstrained utility of the generated synthetic datasets in practice : synthetic datasets bring no privacy issues to subscribers, and may be used by any entity willing to perform realistic network simulations.

Our study is performed on an anonymized dataset collected at the core of a major 3G network of Mexico's capital (Section 2). 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 SMS 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. Moreover, the dataset provides information about age and gender for more than half million subscribers.

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, age, and gender (Section 2). 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. Then, in order to be able to consistently analyse the usage heterogenity of a larger number of subscribers, we classify them into six distinct profiles according to their usage pattern (Section

3). We finally model the usage pattern of these six subscriber profiles according to two different journey periods : peak and non-peak hours. Using a sample and numerous statistical tools, we show the effectiveness of our traffic modeling, which is capable of consistently imitating different subscribers profiles in two journey periods, when compared to the original traffic dataset (Section 4). Our main outcome is *a synthetic measurement-based mobile data traffic generator, capable of imitating traffic-related activity patterns of six different categories of subscribers, during two time periods of a routinary normal day in their lives.* We discuss implications of our contributions in Section 5 and related work in Section 6. Finally, Section 7 concludes this paper. In this paper, user and subscriber will be used interchangeably.

2 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 highlight the concept of a 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 [9]. 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.

Furthermore, due to a special characteristic of this dataset, information about age and gender is available for 548,000 subscribers. This allows us to investigate the interesting relation between users' age, gender, and network traffic demands, which can be used by telecommunication operators to better set subscription plans.

Due to the routinary behavior of people [2] and the large scale dataset, it suffices to study a subset of the whole dataset in order to capture the daily behavior 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' behavior. 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 data from midnight to 1am of all days in the following analysis. This does not affect our methodology since it 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' behavior 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 to model detailed and realistic data traffic demands over time.

In the following, we study the behavior of mobile subscribers in terms of traffic they generate. The analysis are performed according to four main traffic parameters : number of sessions, inter-

arrival time (refered as IAT, the difference between the arrival timestamps of subsequent sessions of the same subscriber), session duration, and volume of traffic.

2.1 Traffic dynamics

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 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 [10]). In the studied week, this average difference is 5%.

Fig. 1(c) shows the CDF (Cumulative Distribution Function) 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 [10].

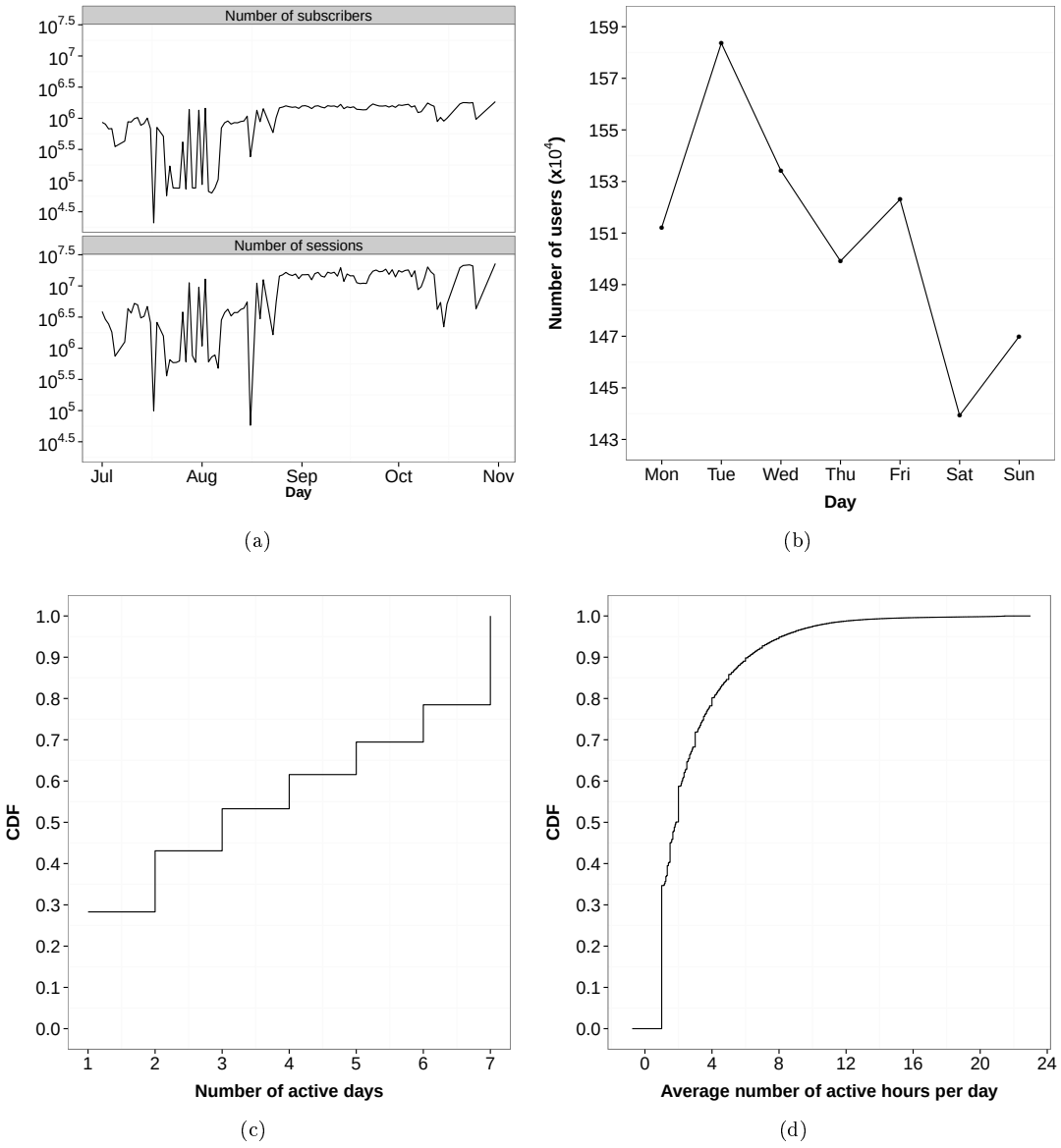
Similarly, in Fig. 1(d), we show the CDF of the average number of active hours of the subscribers per day. We see that *most of subscribers generate traffic on few hours during the day.* Indeed, on an average 80% of the subscribers generate traffic for up to 4 hours each day. If we consider a longer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

Fig. 2(d) shows the total number of sessions per user per day of the week. *There is a slightly less amount of sessions per user during weekends and a general similarity between the cumulative values for all days.* For instance, considering users with up to 10 sessions per day, the difference between the number of sessions per user on weekdays and weekends is 4%, and 0.1% considering up to 100 sessions per day.

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 duration length of sessions.* 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 [10]).

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 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* [11] 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 laying 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 session, we can



(a) (b)

(c) (d)

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.

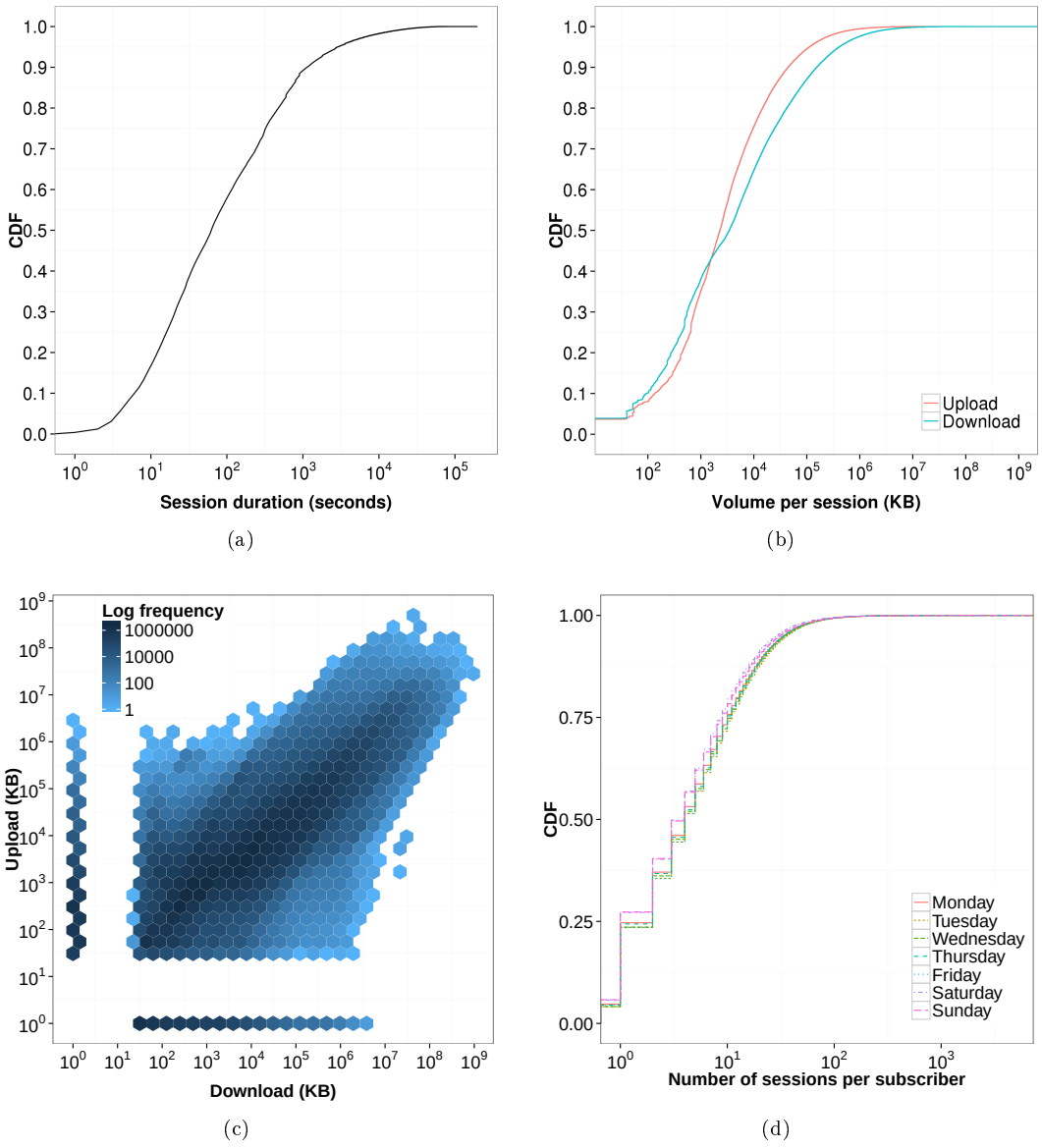


FIGURE 2 – (a) CDF of session duration in seconds per subscriber during the week. (b) CDF and (c) bin plot of the upload and download volume during the week. (d) Number of session per subscriber per day of the week.

expect the uploaded volume to be higher as well. Indeed, the Spearmans's correlation coefficient between per-session upload and download traffic is 88%. 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.

2.2 Temporal dynamics

It is common knowledge that some hours tend to be more active than others when it comes to users routinary daily activities. In this context, peak hours present high frequency of requests and volume of traffic, while 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 behavior 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 behavior 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 results show 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. *Similar to the number of sessions behavior (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) of subsequent sessions of the same subscriber. The high IAT shown from 4am to 8am is a complementary behavior to the low average number of sessions on the same hours present in Fig. 3(a). This is expected and 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 day-wise similarity on number of sessions, volume of traffic, and inter-arrival time traffic parameters. 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' behavior using the Relative Standard Deviation (RSD). RSD is the absolute value of the 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 calculated using the values of the parameters of the same hours for all the days, e.g., the RSD for the number of sessions at 10 a.m. among all days is 2.08%. It is possible to see that the maximum variability is small for all parameters : 3.4% for number of sessions, 1.9%

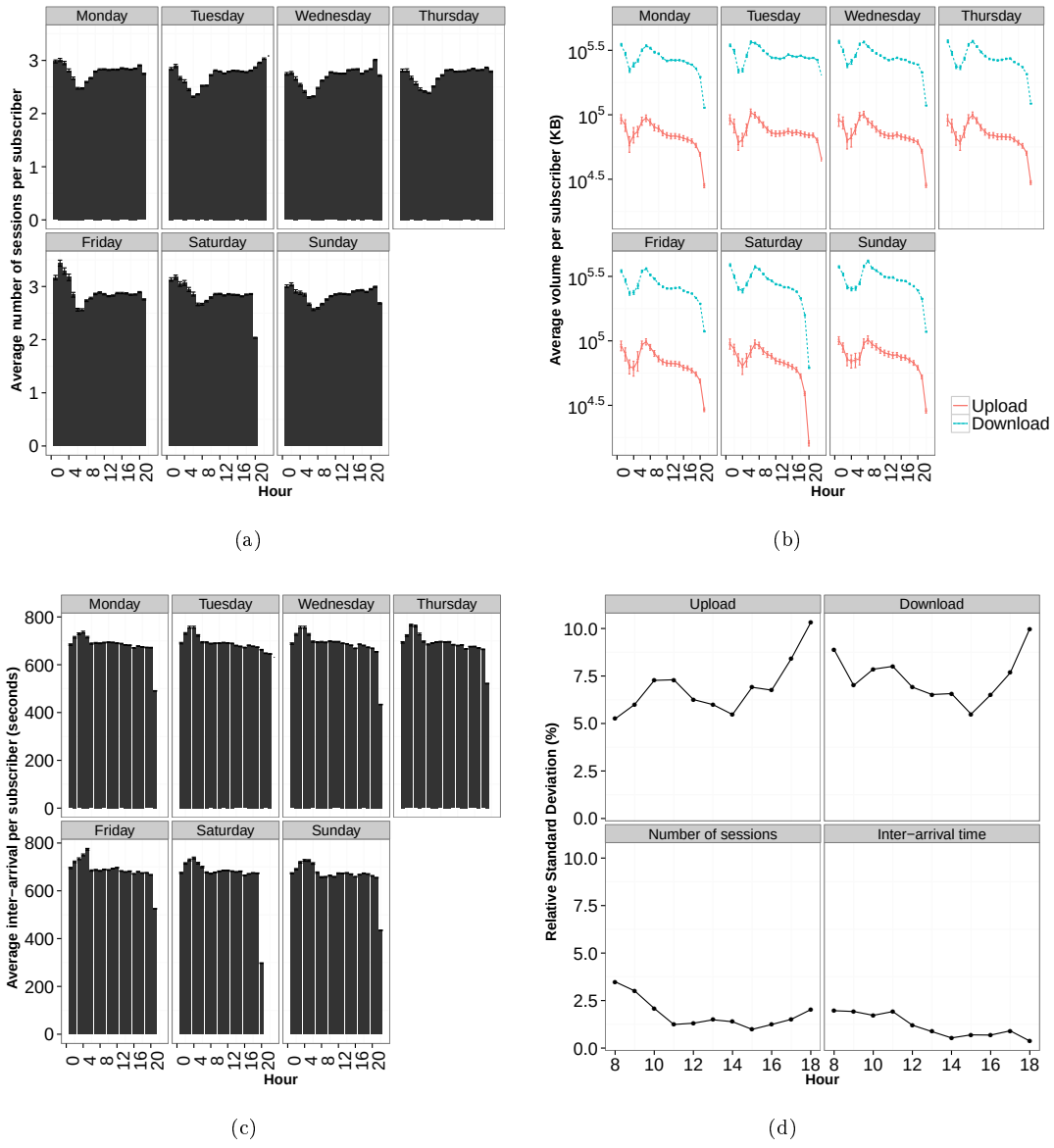


FIGURE 3 – (a) Average number of sessions per user during the week. (b) Volume of traffic for upload and download during the week. (c) Inter-arrival time per subscriber during the week. (d) Relative Standard Deviation per parameter.

for inter-arrival time, 10.3% and 9.9% for upload and download volumes, respectively. In order to show that the variability within the day is higher than the variability among the same hours on different days, we have calculated the maximum RSD of each parameter on all hours of each day. For instance, the variability for the uploaded volume on all hours on Friday is 12.7%. The results shows that, on average, 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.*

Contrarily to our findings, previous related studies considering phone records (or CDRs) [3] show that subscribers behavior in terms of call traffic have strong variation with time and day of the week. Instead, our results show the consideration of real data traffic (instead of call traffic) (1) reveal a different facet of subscribers behavior and (2) stress the imprecisions brought by CDRs analysis to the resource allocation planning.

The similarity of the temporal activity patterns among different days of the week is due to people's natural routinary behavior. Therefore, *we select one day (28th August 2013, a Wednesday) of the week to perform our extensive per-hour analysis and distinguish users profiles.*

2.3 Age and gender dynamics

Among the 2.8 million subscribers in the week mentioned in Section 2.1, a subset of 548 thousand of them present personal information regarding age and gender. All analysis in this section refer to this subset of users. Thus, to better understand how age and gender impacts traffic demands, hereafter, we present our analysis on the traffic parameters when considering these new social information.

As any study considering social aspects of participating entities, it is important to understand in which cultural context the measurements are made. Similarly to many Latin American countries, Mexican culture presents gender wage gap that disfavours women [12]. Consequently, having less purchasing power the Mexican women consume less goods. As a probable consequence, from almost half million users of the considered dataset, 56% are men and 44% are women.

Fig. 4(a) depicts the population pyramid grouped by age and gender. This graphic shows the frequency of age and genders' occurrences with females shown on the left and males on the right. *Regardless of the gender, it is possible to see a higher number of subscribers with age range from 25 to 34 years old.* Indeed, 33% of the subscribers fall in this range.

To ease the understanding of the per-age behavior, we have defined 4 age ranges : $[15, 24]$, $[25, 34]$, $[35, 49]$ and $[50, 85]$, i.e., users younger than 25, from 25 to 34, from 35 to 49, and over 50 years old. Users younger than 15 and older than 85 years old were removed from the trace. In effect, the small amount of users in those two groups make it difficult to draw any statistical conclusion about them. Fig. 4(b) shows the percentage of subscribers grouped by gender and age ranges. It is possible to see a higher percentual of male (and consequently less female) users in all age ranges. An interesting aspect of this graphic is the increasing gap between the genders as the age range progresses. To undercover this aspect we have plotted Fig. 4(c). It shows the percentage of users per age and gender. It is interesting to see that the gap increases with increasing age. The Spearman's correlation between age and age percentage per gender is 87% per male and, consequently, -87% per female, i.e., in our dataset the *male participation percentually increases as the user age increases. Conversely, the female participation decreases with the increase of the age.*

Fig. 4(d) shows the percentage of active users per age and day. An interesting aspect in this graphic is shown for Saturday and Sunday, that have different age range activities when compared to the rest of the days of the week : The absence of the gap present on weekdays from

4 am to 8 am for users within the [25, 34] range, i.e., an activity growth for subscribers from 25 to 34 years old. This is probably due to the nightly activities that usually attracts younger people on weekends, e.g., bars and night clubs.

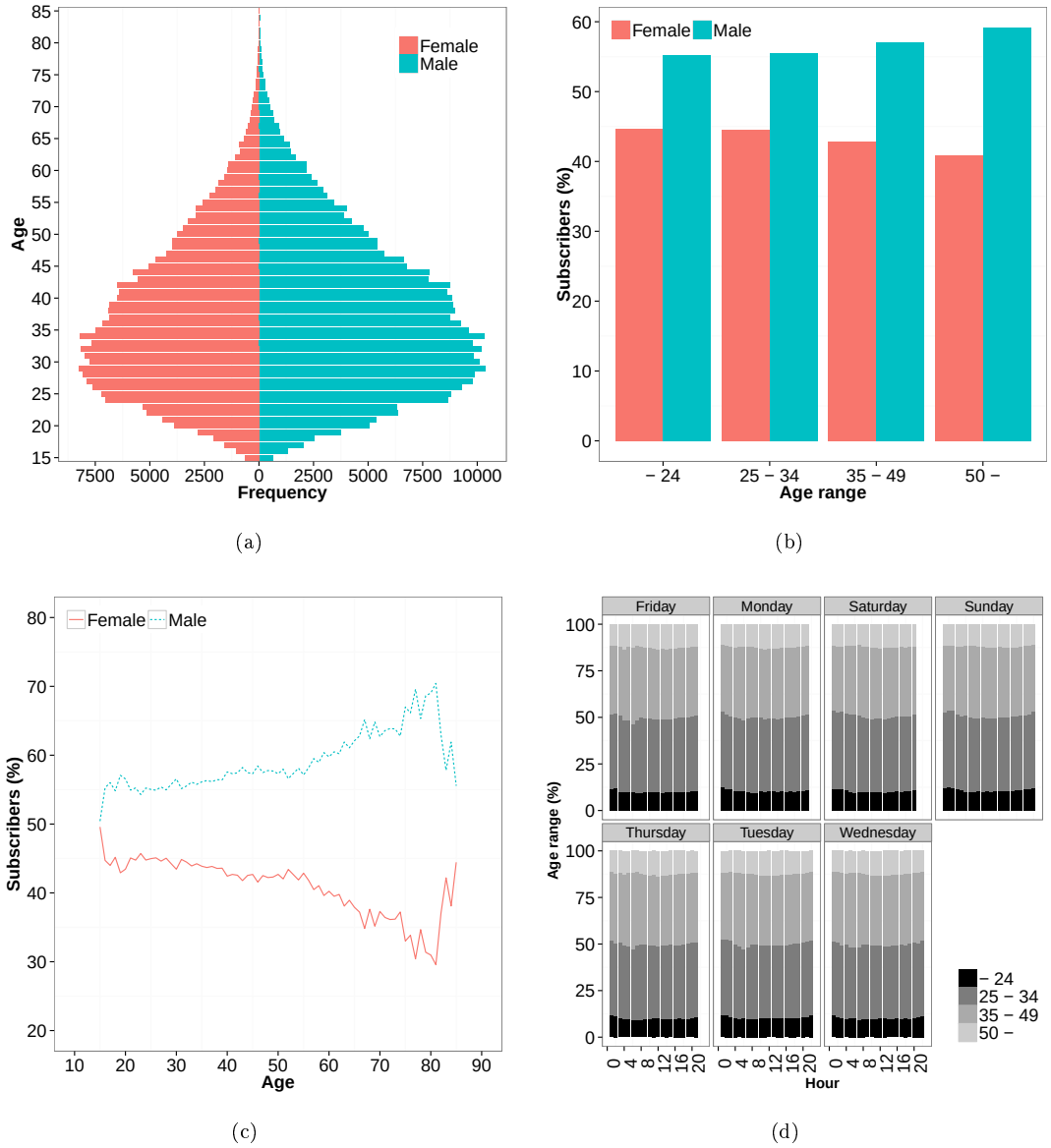
Fig. 5(a) shows the frequency of the number of sessions performed by the subscribers grouped by their ages in all days of the week. In order to improve its visualisation, it does not display the few occurrences in which the number of sessions surpassed 500. Still, it depicts 99.99% of the data related to subscribers' number of sessions. Similar to the day-wise similarity presented in Section 2.2, this graphic shows that the age-wise number of sessions is similar on different days of the studied week. Regardless of the age, most of users present low and similar number of sessions per day (see Fig. 2(d)). Briefly, *per age behavior shows that younger subscribers tend to have peak number of sessions that are higher than older subscribers.*

Fig. 5(b) better shows the decreasing behavior of the traffic parameters with the increase of the age regardless of the gender. It depicts the mean of four traffic parameters by user grouped per age and gender. As there are few users older than 70 years old, their mean values tends to be noisy. If we consider users up to 70 years old, there is a high negative correlation between age and each of traffic parameters for males and females, respectively, -96% and -95% with volume of traffic, -85% and -71% with number of sessions and -63% and -78% with session duration. It means that as the age grows, the value of each of those traffic parameters decrease. Except from the inter-arrival time, there is a clear gap between the maximum and the minimum values for each of the parameters from younger to older subscribers, mainly regarding the total volume of traffic. In order to measure this difference, we have calculated the fraction of the traffic parameters from the oldest age range divided by the youngest one. Indeed, *users from the youngest age range generate, on average, 52% more traffic volume, 21% more sessions, 12% longer sessions with the same inter-arrival time.* Generally speaking, in our dataset *users' network activity tend to decrease with the increase of their age.* Our analysis also show the same decreasing activity when subscribers are grouped by their genders, i.e., *it is related to the age of the subscribers and not a behavior of a specific gender.*

Fig. 5(c) and 5(d) show the CDF of number of sessions and CDF of session duration, respectively, grouped by age range and subscribers' gender. As already discussed, the mean network demands is higher for younger users than for older users. Grouping users by age range diminishes this gap when compared to the per-age analysis, but allows us to see the cumulative differences. For both genders, (1) 80% of the subscribers of the oldest age range and 76% of the youngest age range generate up to 10 sessions during the day and (2) 48% of the subscribers of the oldest age range and 43% of the youngest age range generate sessions up to 15 minutes during the day. *In summary, our analysis shows that similar number of sessions and session duration results are seen when users are grouped by age range, irrespective of the subscribers gender.*

3 Subscriber profiling methodology

Although having their own repeated routine, human behavior in terms of content demand is highly heterogenous, as many other human activities. While some subscribers rarely generate mobile data traffic, others demand a few or even a large amount of gigabytes each day. To analyse such different levels of activity, we group subscribers into a limited number of profiles. The profiles are defined according to two traffic parameters, i.e., traffic demands (i.e., volume of traffic) and activity behavior (i.e., number of sessions). Such parameters are extracted from a sample set of the considered dataset describing subscribers' traffic demands. The profile definition is performed in three phases. First, the similarity metric between all pairs of subscribers on a subscribers' sample set is measured according to the two traffic parameters. Second, subscribers



(a) (b)

(c) (d)

FIGURE 4 – (a) Population pyramid grouped by age and gender. (b) Subscribers by gender per age ranges. (c) Percentage of active users by age. (d) Percentage of active users by age range.

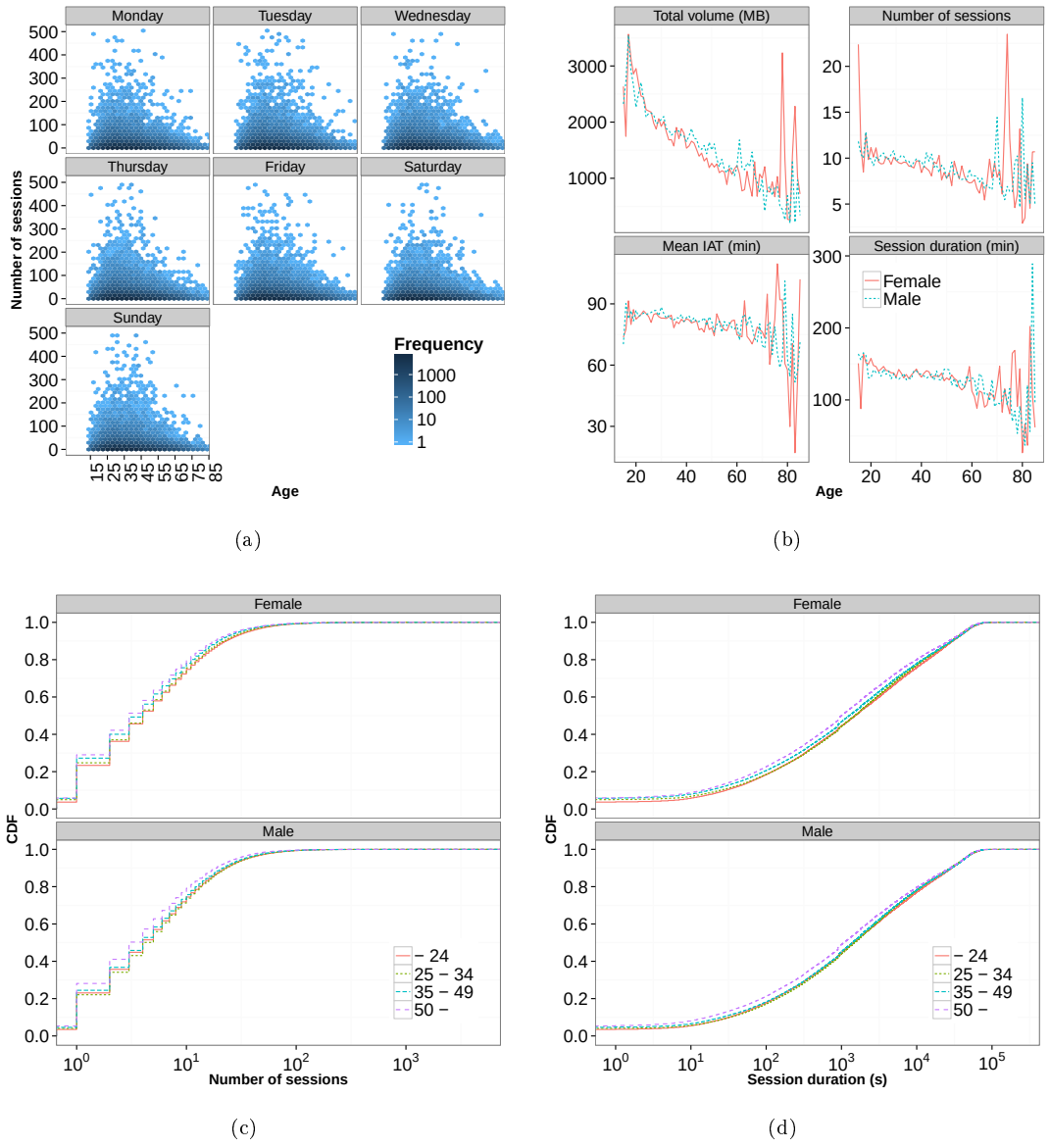


FIGURE 5 – (a) Frequency of sessions per age and day. (b) Mean metrics per age and gender. (c) CDF of the number of sessions per age range and gender. (d) CDF of the session duration per age and gender.

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 previously 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.

3.1 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 [13]. 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 a similarity comparison according to *volume of traffic* and *number of sessions*. These traffic parameters allow us to make a comparison between two different subscribers behavior and will be considered at the clustering and classification procedures (discussed in the next section).

Each subscriber i can be effectively represented by the 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 grained representation of a subscriber is costly in terms of 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

$$V_t^i = \sum_{k \in \tau_t^i} v_k^i. \tag{1}$$

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

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

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.,

$$w_{ij}^{\vartheta} = \|\vartheta^i - \vartheta^j\|. \tag{3}$$

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

$$w_{ij}^{\eta} = \|\eta^i - \eta^j\|. \tag{4}$$

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 $\mathbb{P}$. The remaining users (i.e., $\mathbb{S} - \mathbb{S}'$) are then classified into the previous defined clusters.

3.2 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 dendogram structure [14]. The hierarchical clustering algorithm we choose is the *Average Linkage clustering method*, also known as *Unweighted Pair Group Method with Arithmetic Mean (UPGMA)* [14].

Recall we first group a sample set of $|\mathbb{S}'|$ subscribers into $|\mathbb{P}|$ clusters. Then, we classify the remaining $|\mathbb{S} - \mathbb{S}'|$ subscribers into $\mathbb{P}$. Thus, UPGMA starts by first considering each vertex of the given graph $G(\mathbb{S}', \mathbb{E})$ 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. In our context, it merges together the two clusters that are more similar in terms of 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 dendogram), 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 [15]. Contrarily to related works that have implemented and applied very few of them [10] and in order to avoid to be biased by a specific criteria, we have implemented and used 23 stopping rules, namely Ball-Hall, Beale, Cubic Clustering Criterion, Calinski-Harabasz, C-index, DB, Duda, Dunn, Frey, Friedman, Hartigan, Krzanowski-Lai, Marriot, McClain-Rao, Pseudot2, Ratkowsky-Lance, Rubin, Scott-Symons, SDbw, SD, Silhouette, TraceW, TraceCovW [15, 16, 17, 13, 18, 19].

For the sake of illustration, we will briefly describe the C-Index [15] 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 to 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 value is the one that grants the best separation among clusters.

Fig. 6(a) shows the C-Index index values as a function of the number of clusters, when number of sessions similarity is considered at the distance compution between pairs of users. C-Index considers choosing the best number of clusters based on its minimum index value. Thus, the best number of clusters is 2 according to Fig. 6(a).

Similarly, each other 21 implemented stopping rules listed above define their best number of clusters to be used. In Fig. 6(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 3 as the best profiles, when clustering subscribers by their traffic volumes.

Profiling occurs then in four stages : (1) building a similarity graph with $|\mathbb{S}'|$ 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 $|\mathbb{S} - \mathbb{S}'|$ 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

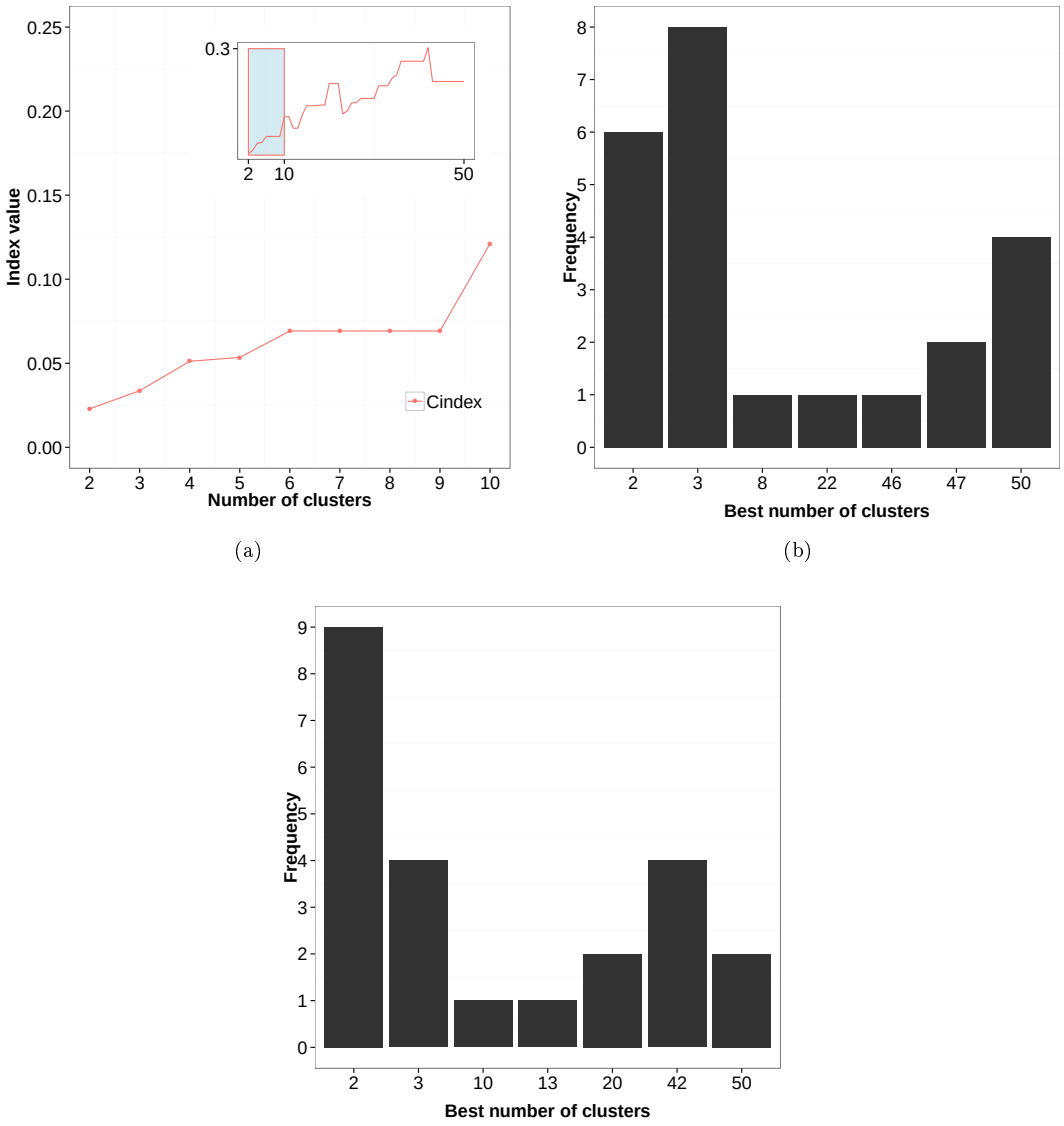


FIGURE 6 – (a) C-Index values and respective number of clusters when re-clustering subscribers at the 3rd defined "traffic-volume"-based cluster, according to the number of sessions similarity. (b) Histogram of best number of "traffic-volume"-based clusters indicated by the assessed stopping rules. (c) Histogram of best number of "number of sessions"-based clusters indicated, when re-clustering subscribers at the 2nd defined "traffic-volume"-based cluster.

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 the results shown in Fig. 6(b), $|\mathbb{P}| = 3$ weighted subgraphs $\{G_1(\mathbb{S}'_1, \mathbb{E}), G_2(\mathbb{S}'_2, \mathbb{E})\}, G_3(\mathbb{S}'_3, \mathbb{E})\}$ are created. At the end of the first round, the final classification of $|\mathbb{S} - \mathbb{S}'|$ subscribers takes place. The next execution round initiates with a new hierarchical clustering being performed inside each initially defined "traffic volume"-based cluster. This time G_1, G_2 and G_3 are weighted according to the *number of sessions similarity* (Eq. (4)). Finally, for each of these three initial clusters, two "number of sessions"-based clusters are defined after the second round of stopping rules execution (e.g., Fig. 6(c)), totalizing six subscribers profiles. Due to space constraints, we will not show all stopping rules results. The second round ends with the classification of the remaining $|\mathbb{S} - \mathbb{S}'|$ subscribers into the six defined profiles. Next section better details our subscriber profiling.

3.3 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, $|\mathbb{S}'| = 10000$ to be used in the clustering procedure). 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 behavior 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' behavior 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. Nevertheless, the optimal size of the time slot is still an open problem [20].

Our profiling methodology resulted in *six profiles*, and we have named them as follows : Light Occasional (LO), Light Frequent (LF), Medium Occasional (MO), Medium Frequent (MF), Heavy Occasional (HO) and Heavy Frequent (HF). *Light* profiles contain subscribers that generate up to 17 MB of data during the day, *Medium* profiles have subscribers that generate between 17 MB and 560 MB of traffic during the day, and *Heavy* profiles contain users that generate more than 560 MB of traffic during the day. Likewise, *Occasional* profiles contain subscribers that generate less connection sessions, whereas *Frequent* profiles contain users generating more connections per day. Tables 1, 2, and 3 show the characteristics of each of the profiles.

TABLE 1 – Characteristics of the Light profile

	Light	
Volume	29 KB to 17305 KB ($\approx$ 17 MB)	
N° of subscribers	418843	
	Occasional	Frequent
N° of sessions	1 to 10	11 to 224
N° of subscribers	405848	12995

TABLE 2 – Characteristics of the Medium profile

	Medium	
Volume	17306 KB to 560044 KB (≈ 560 MB)	
N° of subscribers	610917	
	Occasional	Frequent
N° of sessions	1 to 51	52 to 1926
N° of subscribers	598340	12577

TABLE 3 – Characteristics of the Heavy profile

	Heavy	
Volume	560046 KB to 655769309 KB (≈ 650 GB)	
N° of subscribers	487141	
	Occasional	Frequent
N° of sessions	1 to 316	317 to 8737
N° of subscribers	484959	2182

In Fig. 7, we show the dynamics of the traffic parameters per subscribers' class per hour. Fig. 7(a), 7(b), and 7(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 :

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

where τ_t^i is the set of all sessions of subscriber i 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 $\zeta^i = \sum_{t \in D} IAT_i^t$.

From Fig. 7, we can see that our methodology well separates the profiles, i.e., the *occasional* and *frequent* subscribers have their values clearly separated. Note that an aggregated traffic analysis would not allow us to identify and consequently, to imitate the behavior of very light users. In fact, the traffic generated by very heavy users (representing a very small percentage of users in the dataset) would bias the analysis and the synthetic traffic generation.

For each curve in Fig. 7(a), 7(b), and 7(d), 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*.

3.4 Profile's age and gender

In this section, each of the resulting profiles is assessed by the age and gender of their members. The profiled day D has 1.5 million users, from which 107 thousand have information regarding age and gender. The results shown in this section refer to this subset that counts with 57.6% of male and 42.4% of female users. This subset is consistent with the distribution of users with

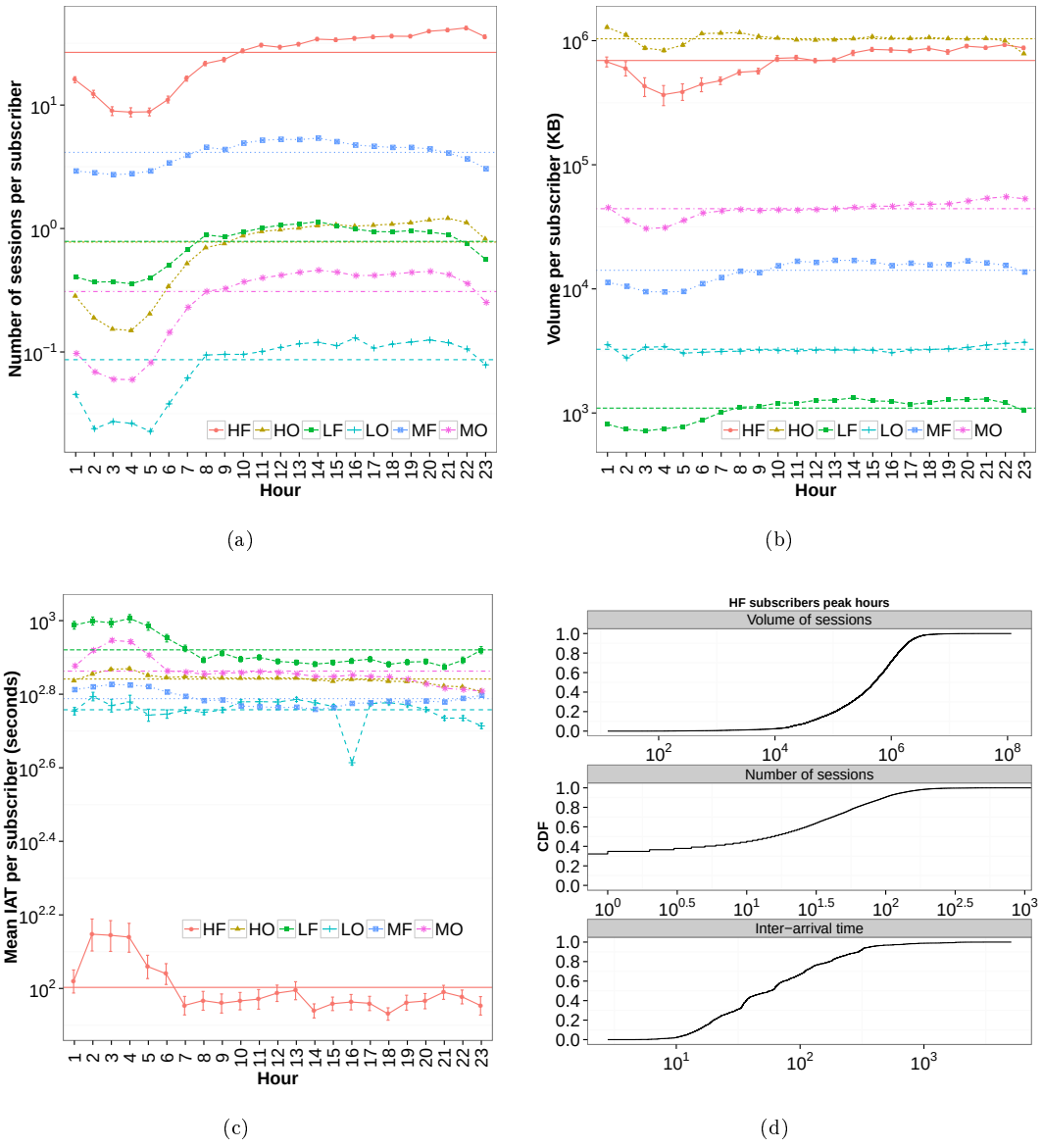


FIGURE 7 – (a) Mean inter-arrival per class. (b) Number of sessions per class. (c) Volume of traffic per class. (d) Empirical CDFs of HF users in peak hours.

available age and gender prior to the profiling process, which counted 548 thousand subscribers over a week (Section 2.3). To evaluate this consistency, we calculate the percentage of users per age on the 548 thousand non profiled users and on the 107 thousand profiled users. Fig 8(a) shows this percentage for each of them. There is a visual similarity between the shape of the two curves as they are strongly correlated, with 99% Spearman's correlation.

Fig. 8(b) shows the percentage of male and female subscribers per class, after the profiling of 107 thousand subscribers. *Most of the classes present higher percentual of male than female, except HF in which female have 1% more users than male.* On average, Light and Medium profiles have 15% more males than females, while Heavy profiles have 6% more male than female.

Fig. 8(c) shows the average subscribers' ages per gender and classes. Due to the large overlapping presented by the confidence intervals (95%), we can assert that *the per-class ages are not significantly different.* That is interesting because *it indicates that the profiles group together users from a wide spectrum of different ages.*

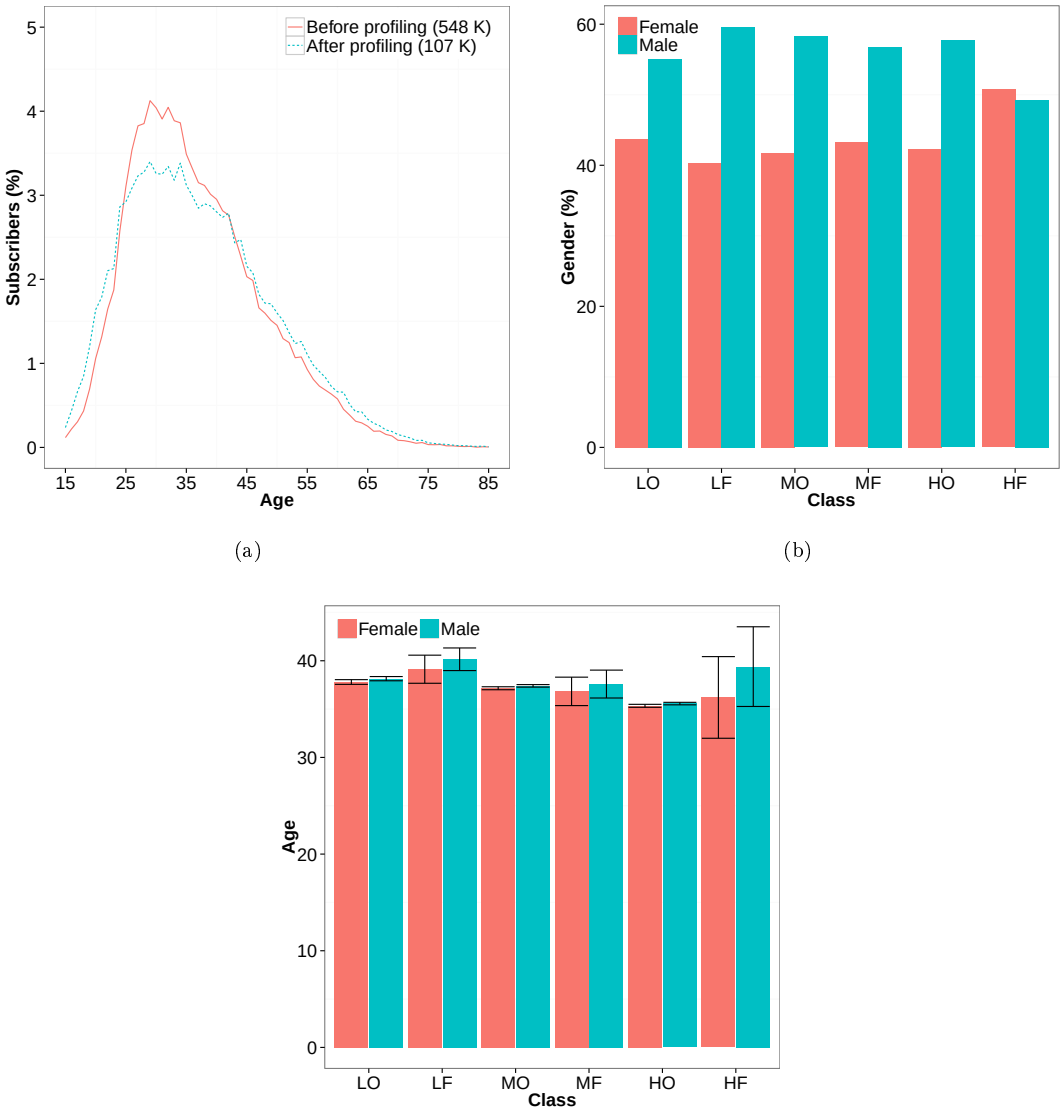
Fig. 9(a) and 9(c) show the CDFs of number of sessions per subscriber per class. The former groups subscribers per age range and the latter per gender. An interesting difference between Occasional and Frequent users is steepness of the CDF curves. *Number of session from Occasional profiles is more uniformly distributed than from Frequent users, which has a very steep slope. It means that most of the Frequent users generate the lowest amount of sessions within the range of their profiles* (recall that the ranges are specified in Table **??**). *For all classes, male users generate, on average and median, more sessions than females.* On Occasional classes the difference is 1% at most, while on Frequent classes the difference ranges from 2% to 19%. The cumulative values show the same results, for instance the third quartile is at most 1% higher for male than female on all Occasional and LF profiles. Moreover, it is 10% higher on MF and HF profiles.

Fig. 9(b) depicts the CDFs of session duration per subscribers' class and age range. On average, profiles do not present statistically different session duration values for each of the age ranges. For instance, the per-class confidence intervals (95%) for each of the age ranges overlap each other by the mean. It means, *the session duration behavior within each of the profiles for a certain age range is not statistically different from the behavior of another age within the same profile.*

Fig. 9(d) presents the kernel density estimation (KDE) curves for the volume per user per gender and class. *There is a similar behavior for male and female subscribers for all the profiles, except HF.* HF male subscribers density curve is narrower than the female one and present a peak around 10 GB. On the other hand, HF female subscribers curve is wider. It means that, among the heavy and frequent subscribers, male present less diverse session volumes when compared to female.

4 Measurement-driven traffic modeling

Realistic network simulations requires a traffic generator capable of imitating actual daily subscribers traffic demands, i.e., has to be consistent with the observations made about the real subscribers in the previous section. Recall that subscribers belonging to different profiles (LO, LF, MO, MF, HO, and HF) have their own specificities in terms of *when* the sessions are generated during the day, and the *volume* generated during each session. Furthermore, each profile of subscribers have different behavior 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.



(a) (b)

(c)

FIGURE 8 – (a) Percentage of subscribers per age before and after profiling. (b) Percentage of subscribers per gender and class. (c) Average subscribers' age per gender and class.

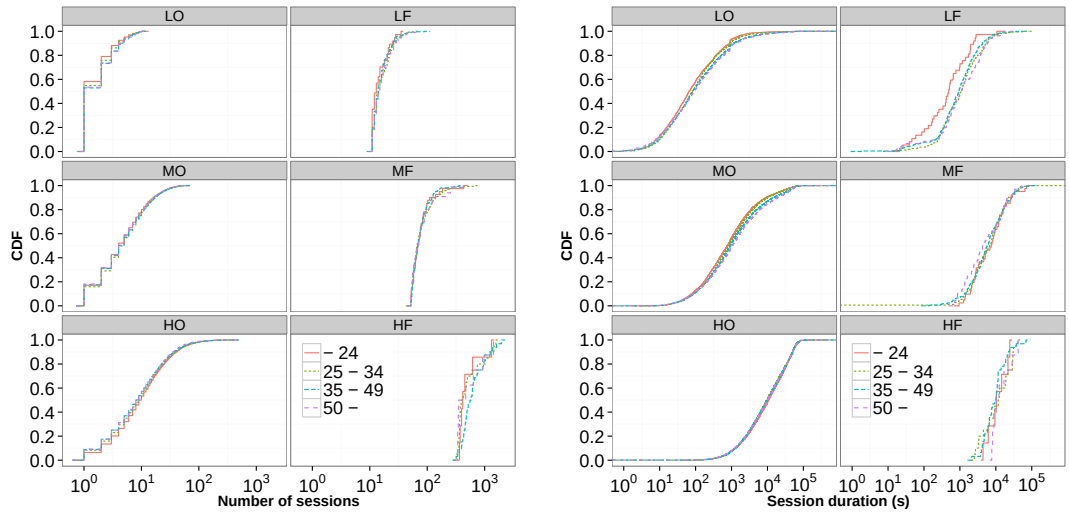


FIGURE 9 – (a) CDFs of number of sessions and (b) session duration per subscribers' class and age range. (c) CDFs of number of sessions and (b) session volume per subscribers' class and gender.

4.1 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 (e.g., Fig. 7(d)) : 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 a 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. *This function will be used 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 [21] 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 [22] 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 4, 5, and 6 list the best fitted distribution functions along with their parameters for all possible combinations of profile and hour-type pair, for number of sessions, traffic volume and inter-arrival time parameters, respectively. 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.

4.2 Synthetic subscriber generation

Generating a synthetic subscriber will first require us to generate a profile type (LO, LF, MO, MF, HO, or HF) for the subscriber. Profile types are assigned randomly, based on the distribution of profiles population observed in the real data. For instance, from Table **??**, we see that 26.7% of the subscribers belong to LO profile, and thus with probability $q_{LO} = 0.267$ we assign LO profile to a synthetic user. Similarly, the probabilities of other profiles are : $q_{LF} = 0.0085$, $q_{MO} = 0.394$, $q_{MF} = 0.0082$, $q_{HO} = 0.319$, and $q_{HF} = 0.001$. We will refer to $q = (q_{LO}, q_{LF}, q_{MO}, q_{MF}, q_{HO}, q_{HF})$ as the *profile pmf*, or probability mass function.

We now briefly describe our procedure for generating a synthetic subscriber (for a detailed algorithm, refer to [23]). *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. The result of the generation is a list of sessions per user. 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, an average session volume V_t^i, a mean inter-arrival time IAT_t^i from the appropriate distributions (i.e., the fitted distribution corresponding to the profile and hour-type pair) listed in Tables 4, 5, and 6, respectively. The volume per session v_k^i

(for $k \in \tau_t^i$, see Section 4) 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. By varying t over the 24 hours in a day, we obtain a synthetic subscriber traffic for one day.

TABLE 4 – Number of sessions : distributions and parameters

Number of sessions			
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Neg-binomial	$n = 0.1139$, $p = 0.09$
	HF		$n = 0.4703$, $p = 0.01$
	MO		$n = 0.1772$, $p = 0.3$
	MF		$n = 0.7588$, $p = 0.13$
	LO		$n = 0.1885$, $p = 0.62$
	LF		$n = 0.4802$, $p = 0.32$
Non-Peak	HO	Neg-binomial	$n = 0.0448$, $p = 0.1$
	HF		$n = 0.1437$, $p = 0.01$
	MO		$n = 0.0536$, $p = 0.3$
	MF		$n = 0.3146$, $p = 0.08$
	LO		$n = 0.0810$, $p = 0.66$
	LF		$n = 0.2405$, $p = 0.33$

TABLE 5 – Session volume : distributions and parameters

Session volume			
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Weibull	$k = 0.49$, $\lambda = 476551.7$, $x_0 = 30$
	HF		$k = 0.81$, $\lambda = 774639.6$, $x_0 = 40$
	MO		$k = 0.59$, $\lambda = 31936.8$, $x_0 = 29$
	MF		$k = 0.80$, $\lambda = 13959.4$, $x_0 = 37$
	LO		$k = 0.85$, $\lambda = 3228.7$, $x_0 = 29$
	LF		$k = 0.92$, $\lambda = 1181.7$, $x_0 = 33$
Non-Peak	HO	Weibull	$k = 0.50$, $\lambda = 452332.8$, $x_0 = 30$
	HF		$k = 0.63$, $\lambda = 384935.6$, $x_0 = 40$
	MO		$k = 0.58$, $\lambda = 26617.7$, $x_0 = 30$
	MF		$k = 0.79$, $\lambda = 10657.9$, $x_0 = 33$
	LO		$k = 0.79$, $\lambda = 2800.1$, $x_0 = 29$
	LF		$k = 1.03$, $\lambda = 873.5$, $x_0 = 34$

4.3 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

TABLE 6 – Session mean inter-arrival times : distributions and parameters

Session mean inter-arrival time			
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Gamma	$\alpha = 1.2517$, $\beta = 0.0017$, $x_0 = 0.5$
	HF	Log-normal	$\sigma = 4.0917$, $\mu = 1.1285$, $x_0 = 4.68$
	MO	Gamma	$\alpha = 1.2990$, $\beta = 0.0016$, $x_0 = 0.5$
	MF	Gamma	$\alpha = 2.2081$, $\beta = 0.0034$, $x_0 = 1$
	LO	Weibull	$k = 0.8508$, $\lambda = 548.24$, $x_0 = 1$
	LF	Gamma	$\alpha = 1.7929$, $\beta = 0.0019$, $x_0 = 2$
Non-Peak	HO	Gamma	$\alpha = 1.2044$, $\beta = 0.0017$, $x_0 = 0.5$
	HF	Log-normal	$\sigma = 3.9374$, $\mu = 0.9822$, $x_0 = 3$
	MO	Gamma	$\alpha = 1.1921$, $\beta = 0.0017$, $x_0 = 0.5$
	MF	Gamma	$\alpha = 2.0301$, $\beta = 0.0034$, $x_0 = 1$
	LO	Gamma	$\alpha = 0.7078$, $\beta = 0.0013$, $x_0 = 1$
	LF	Weibull	$k = 1.1988$, $\lambda = 827.96$, $x_0 = 1$

i and at hour t : (1) the volume in KiloBytes generated and (2) the initial timestamp of the session.

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^e_ϑ denote the PDF (Probability Distribution Function) of the total volume generated by a subscriber active in day e in the original trace, formally defined as $p^e_\vartheta(x) = \sum_{i \in e} \mathbb{I}(\vartheta^i = x)/|\{i \in e\}|$. For a visual comparison, Fig. 10(a) depicts the CDFs corresponding to the PDFs p^D_ϑ and $p^{D'}_\vartheta$ of traffic generated in the original day D and synthetic day D'. We can observe an *almost complete overlap of the two CDFs due to high similarity between the real trace and the synthetic trace.*

We then 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 [24]. It quantifies the similarity between two discrete or continuous probability distributions. Let $p(i)$ and $p'(i)$ be two pmfs, i.e., $\sum_{i=1}^{N} p(i) = \sum_{i=1}^{N} p'(i) = 1$. The BH measure is formally defined as $\rho(p, p') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)}$. However, the BH measure is not a distance metric since it does not satisfy all the metric axioms. Therefore, [25] proposes an alternative distance metric based on the BH measure which is formally defined as $d(p, p') = \sqrt{1 - \rho(p, p')}$. Note that, $d(p, p')$ exists for all discrete distributions and it is equal to zero if and only if $p = p'$. We use d in order to measure the similarity between the original dataset and the synthetic dataset.

We first compute $d(p^D_\vartheta, p^{D'}_\vartheta)$, the distance between the total volume distribution of the original day and the synthetic day. Then, we compute $d(p^D_\vartheta, p^e_\vartheta)$, $e \in \mathbb{D}$ but $e \neq D$, the distance between the original day and remaining days in the original trace. We obtain similar distances for p^e_η and p^e_ζ 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. 10(b), we show the $d(p^D_\vartheta, p^e_\vartheta)$ distances (cf. $d(p^D_\eta, p^e_\eta)$ and $d(p^D_\zeta, p^e_\zeta)$). Also shown in Fig. 10(b) (horizontal dashed line) is the $d(p^D_\vartheta, p^{D'}_\vartheta)$ distance (cf. $d(p^D_\eta, p^{D'}_\eta)$ and $d(p^D_\zeta, p^{D'}_\zeta)$). The traffic model evaluation consists then in verifying whether the $d(p^D_\vartheta, p^{D'}_\vartheta)$ is within the confidence interval of the $d(p^D_\vartheta, p^e_\vartheta)$. As can be seen in Fig. 10(b), *for each distribution,*

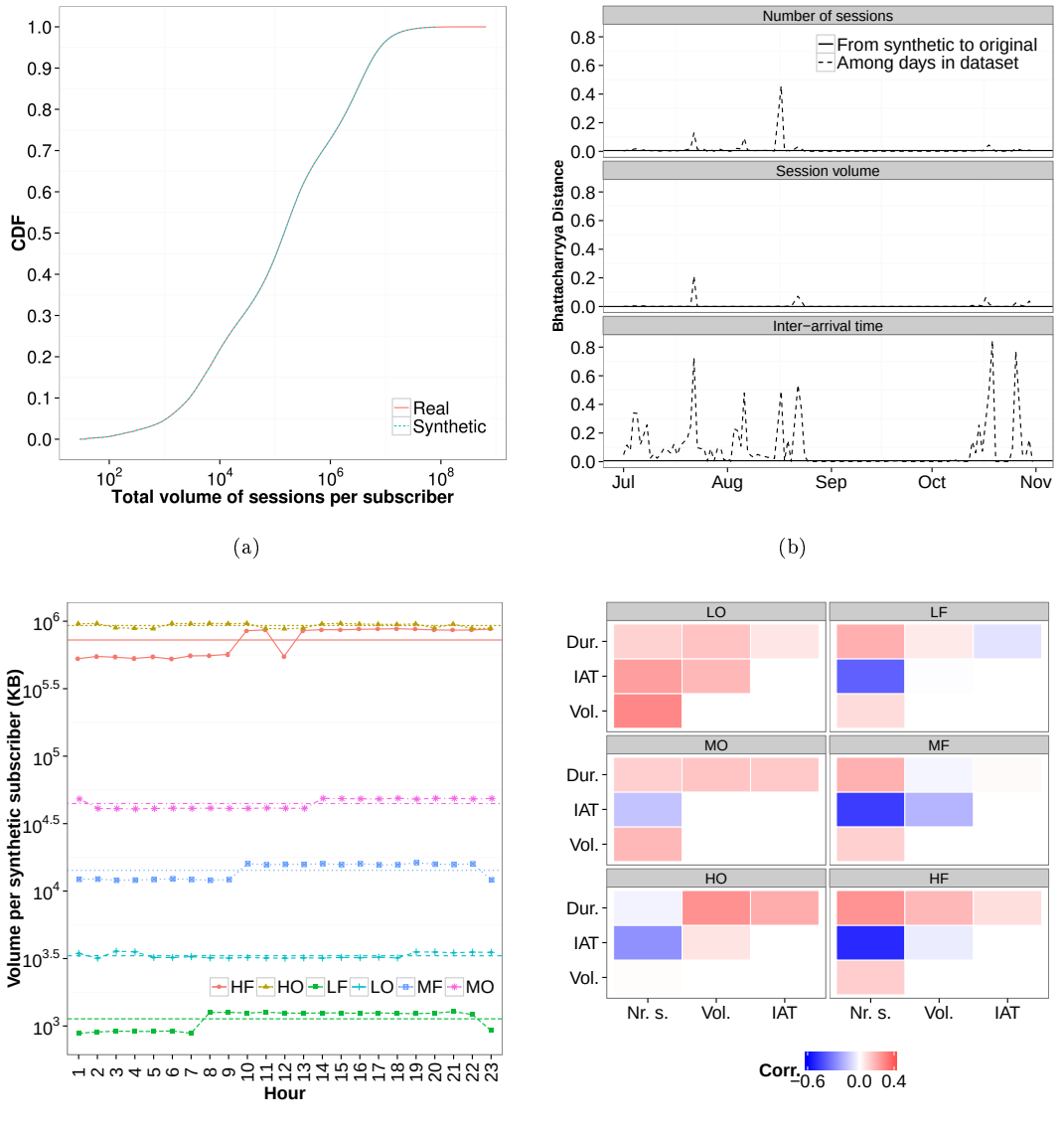


FIGURE 10 – (a) CDF of the total volume generated by real and synthetic subscribers (b) Per-parameter BH distances between original and synthetic trace (dashed line) in D, and between the original trace in D and other days e from the original trace (full line) (c) Volume of traffic per class for synthetic subscribers. (d) Heatmap (better seen in colors) of the correlation between session duration, inter-arrival time and volume of traffic.

the distance of the synthetic day (from the original) is within this confidence interval.

Finally, we applied the profiling methodology described in Section 3 on the synthetic subscribers. By doing so, we classify them and compare the per-class traffic behavior with the one created from the original dataset. Fig. 10(c) depicts the per-class behavior 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. 7(b). For instance, the behavior for peak and non-peak hours is well defined and similar to the one from the original trace.

5 Discussion

In this section, we discuss some issues we judge interesting in the presented work. 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. For example, our dataset shows that the heaviest user generates 22 million times more traffic than the lightest one. Moreover, the 276 thousand lightest subscribers generate similar amount of traffic as generated by a unique heaviest subscriber in the entire day.

Traffic demand is generally described by the set of different traffic parameters that characterize the demands of the users to the network. In this work, we have explored a set of parameters such as inter-arrival time, session duration, number of sessions, and volume of traffic. Alone, each of those parameters were deeply assessed on our previous sections, but it is also interesting to see what is the relation among them.

Fig. 10(d) shows a heatmap (better seen in colors) of the Pearson's correlation between those traffic parameters for all subscribers in all profiles. The intensity of the color on each cell of the matrix indicates how strong is the negative or positive correlation. It is possible to see that the correlation between number of sessions and inter-arrival time goes from a low positive value on LO to a high negative value on HF. Indeed, the correlation between them is 22%, -14% -26% -37% -45% -55% for LO, MO, HO, LF, MF, and HF, respectively. *It means that classes in which subscribers generate more sessions have higher negative correlation with the inter-arrival time. In general, the more sessions a user generates, the shorter they need to be to fit in a certain period of time.* A caveat here is that a user that generates few sessions could generate them in bursts, or sparsely separated in time. The former would result in small IAT and the latter in a larger IAT. For example, a large IAT of one hour is likely to be done for a user with few sessions per day, than a user with, for example, 300 sessions. In the same way, a small IAT could be generated for a user with both high or low number of sessions.

Another important aspect is the relation between volume of traffic and session duration. LO, MO, and HO classes present 13%, 14%, and 26%, respectively, i.e., a growing positive correlation with the session duration. LO, MO, and HO have, on average, 663, 6554 and 18624 seconds of session duration and 5090, 165214 and 6117322 KB of average session volume, respectively. *The growth of those metrics from one Occasional class to the next is due to the necessary increase on the session duration in order to accomodate the volume of traffic, considering that there is no significant raise on the number of sessions from LO, to MO, or HO.*

Finally, it is important to mention the correlation between number of sessions and volume of traffic. The correlation is overall low and positive between these two metrics for all profiles, but its behavior differs completely from Occasional to Frequent users. LO, MO, and HO have 29%, 17% and 0.4% correlation between number of sessions and volume of traffic, respectively, i.e., a decrease from LO, to MO and to HO. It happens because LO users have few sessions and low traffic volume, while MO and HO classes have significantly higher volume of traffic, but still few sessions. Therefore, the correlation is lower for MO and HO than for LO. Differently, LF, MF,

and HF have 9%, 11% and 12% correlation between number of sessions and volume of traffic, respectively, i.e., a growth from LF, to MF and to HF. That is due to HF presenting both high volume and high number of sessions, while LF and MF present lower volume of traffic, but still high number of sessions.

Understanding network demands from users traffic parameters and their correlations is one of the contributions of our work. Moreover, this work provides distributions to model workload characteristics of mobile subscribers' traffic demands and a framework on how to create a traffic generator out of it. Therefore, it has implication in areas related to the design of new applications and network mechanisms as well as network planning such as hotspot deployment [2].

In this latter area, for instance, 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. Besides, a synthetic traffic generator allows the generation of traffic demands of any size of population : While the original traffic dataset allow the sampling of users up to the size of the dataset, a synthetic traffic generator allows to expand this limit.

Another important aspect of the synthetic traffic generator is that it preserves the privacy of the original subscribers from whom the measurements came from. The non-existence of personal data attached to synthetic users allows us to limitlessly share our observations with the community without the necessity of sharing sensitive information inherent of datasets. One may argue that it is possible to anonymize the users identity, but literature shows that many attempts on that direction fail on protecting users privacy [26]. As shown in our analysis, our synthetic users generate traffic consistent with the original dataset and, thus do not carry privacy issues.

6 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.

A significant amount of works in the literature analyze network traffic usage through voice calls and SMS messages, both extracted from traditional Call Detail Records (CDRs). Analysis such as [4, 27, 8, 6, 28, 29, 30] may provide an idea on the activity of mobile network customers but do not describe realistic data traffic demand patterns. In fact, contrarily to data traffic demands, call traffic has the limitation of being sparse in time (i.e., generated only when a voice call or a text message service occurs), which makes cellular users invisible at all other periods of time. Moreover, due to the richness of the data set used in our studies, we can precisely infer traffic activity patterns over time, instead of considering only the times at which users actively generated traffic. This includes the traffic load automatically generated by current smarphone applications (email checks, synchronization, etc). Our analysis also differs from [6, 31], since we target an individual user characterization rather than a network-wide one. Additionally, we aim to profile subscribers by their traffic demands, not by their browsing behavior, i.e., websites they

normally visit as proposed in [32]. Moreover, contrarily to [8], we focus on an activity pattern characterization of a normal day, known to represent typical network usage.

Still, other works such as [7, 3, 33, 34, 35, 36], or [37] have categorized actual mobile traffic usage. For instance, [7, 3, 33, 34], and [37] have only considered total traffic volume when characterizing users' behavior. Studying this metric alone does not reflect the activity variation of users : i.e., number and frequency of requests. [35, 36] 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. 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 profiling of individual users behavior are considered.

With regards to age, gender and network usage investigation, [29, 30] are the most prominent works in the literature. Both studies analyze how gender and age affects the usage of voice calls and text messages, but contrarily to our work, no data traffic analysis is provided.

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 data sessions from about 6.8 million smartphone users. Moreover, we propose a framework that automatically classifies those users according to their traffic demands into a limited number of profiles. Our approach takes advantage of repetitive users behavior due to their daily routines. Furthermore, we provide 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 behavior when compared to original dataset.

As future work, we aim to model sessions' transfer rate and duration. Moreover, we intent to study the existence of real-world aspects on the synthetic trace other than the inter-arrival time, e.g., temporal auto-correlations of each measure. Besides, we envision to create a comparison scenario with a simple baseline model, e.g., that does not separate users by classes, but considers a single user class and the mapping to one single probability distribution from the two traffic parameters. Additionally, we intend to apply and evaluate our traffic generator on different problems such as network planning. Relying on a future availability of geographic data, we plan to study the traffic parameters' spatial correlation.

Table des matières

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] 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.

[5] J. Wortham, "Cellphones now used more for data than for calls," *New York Times*, May 2010.

[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] Alcatel-Lucent, "Alcatel-lucent 9900 wireless network guardian," White Paper, Dec. 2012.

[10] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. of IEEE Infocom*, Apr. 2011.

[11] 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.

[12] International Trade Union Confederation, "Frozen in time : Gender pay gap unchanged for 10 years," Tech. Rep., 2012.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] R. B. D'Agostino and M. A. Stephens, *Goodness-of-Fit-Techniques*. CRC Press, Jun. 1986, vol. 68.

[22] 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.

[23] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modelling in a large metropolitan area," INRIA, Tech. Rep., 2014. [Online]. Available : https://hal.inria.fr/hal-01073129v4/document

[24] 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.

[25] 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.

[26] President's Council of Advisors on Science and Technology, "Big Data and Privacy : A Technological Perspective," Executive Office of the President, Tech. Rep., 5 2014.

[27] 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.

[28] R. A. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, E. Varshavsky, and C. Volinsky, "Clustering anonymized mobile call detail records to find usage groups," Workshop on Pervasive and Urban Applications (PURBA), 2011.

[29] A. Stoica, Z. Smoreda, C. Prieur, and J.-L. Guillaume, "Age, Gender and Communication Networks," in *NetMob 2010 Workshop on the Analysis of Mobile Phone Networks*, V. Blondel and G. Krings, Eds., May 2010.

[30] A. Mehrotra, A. Nguyen, J. Blumenstock, and V. Mohan, "Differences in phone use between men and women : Quantitative evidence from rwanda," in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ser. ICTD '12. New York, NY, USA : ACM, 2012, pp. 297–306. [Online]. Available : http://doi.acm.org/10.1145/2160673.2160710

[31] 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.

[32] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3g network using hourglass co-clustering," in *Proc. of ACM MobiCom*, Sep. 2010.

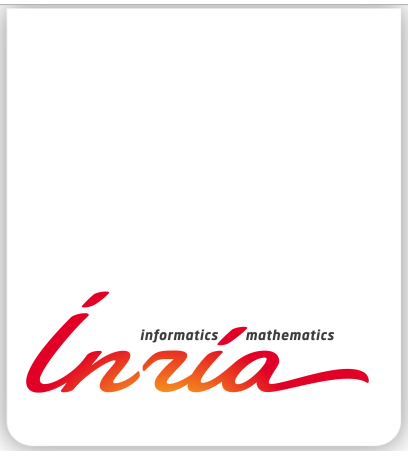
[33] A. Vaccari, L. Liu, A. Biderman, C. Ratti, F. Pereira, J. Oliveirinha, and A. Gerber, "A holistic framework for the study of urban traces and the profiling of urban processes and dynamics," in *Proc. of Int. IEEE Conf. on Intelligent Transportation Systems (ITSC)*, Oct. 2009.

[34] 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.

[35] 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.

[36] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, "Towards estimating the presence of visitors from the aggragate mobile phone network activity they generate," in *Proc. of Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.

[37] Q. Lin, "Mobile customer clustering analysis based on call detail records," in *Communications of the IIMA*, vol. 7, no. 4, 2007.



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