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

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

Measurement-driven Mobile Data Traffic Modeling in a Large Metropolitan Area

Eduardo Mucelli Rezende Oliveira[◇,×], Aline Carneiro Viana[×], K. P. Naveen[×], and Carlos Sarraute[⋆]

[◇] École Polytechnique, France [×] INRIA, France [⋆] Grandata Labs, Argentina

***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 behavior 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.**

I. Introduction

The recent boost of mobile data consumption 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 thus face a number of challenges, including the need to operate in such extreme environments. In this context, *understanding mobile data traffic demands per user* is crucial for the evaluation of data offloading solutions designed to alleviate cellular networks [1], thus favoring 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 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.

Contrarily to most related work in the literature modeling call traffic (commonly 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 traffic provides an intuition of users activity in the network: voice calls and SMS. Since smartphones are now used more for data than for calls [2], the use of calls/SMS for investigating traffic demands is not enough for dimensioning network usages. In addition, besides being sparse in time [3], it does not describe the background traffic load automatically generated by current smartphone applications (e.g., email checks, synchronization).

Our contributions are twofold: First, our analyses provide a *precise characterization of individual subscribers traffic behavior clustered by their usage pattern*, instead of a network-wide data traffic view [4]. Note that, the high variance of individual subscriber behavior (in terms of traffic demands and in time) and the use of large scale datasets make this task complex. Next, we provide *a traffic generator that synthetically, still consistently, reproduces real traffic demands*. A synthetic traffic has positive implications on network resource allocation and planning, on infra-structure testing, or on protocols and services validation. Moreover, the synthetic traffic carries no personal information from the original users, thus it carries no privacy issues unlike the original dataset.

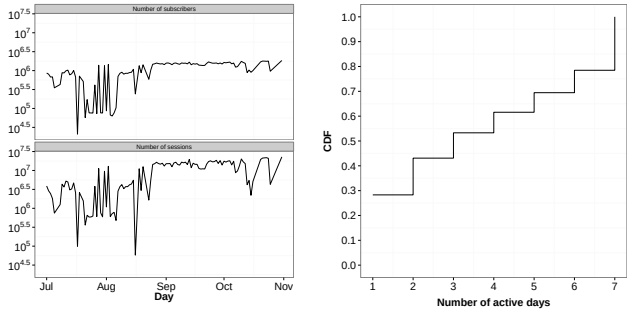
Our study is performed on an anonymized dataset collected at the core of a major 3G network of Mexico's capital, consisting of data traffic associated with 6.8 million subscribers collected over 4 months from July to October 2013. We first analyse subscriber's traffic usage habits as a function of time (Section II). 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 into four distinct profiles according to their usage pattern (Section III). We finally model the usage pattern of these four subscriber profiles according to two different journey periods: peak and non-peak hours. Using a sample test set and numerous statistic 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 IV).

Finally, 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 routine normal day in their lives.* Throughout this paper, the terms user and subscriber will be used interchangeably.

II. Dataset

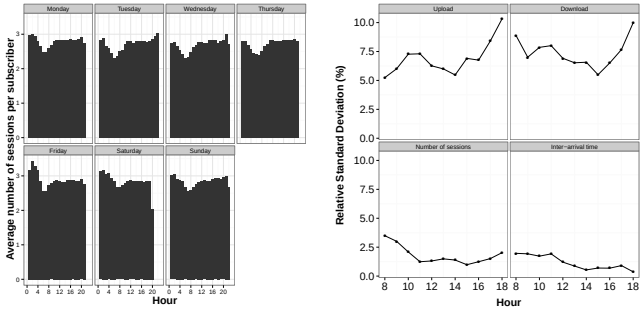
This dataset captures subscribers' traffic activities generated by 6.8 million smartphone devices located within the large urban area of Mexico city. The data includes information about subscribers' *sessions* that took place from 1st July to 31st October, 2013. It is important to highlight the concept of "session" in our work. In the 3G standards 3GPP or 3GPP2, a session is created when the radio channel is allocated to a

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



(a) (b)

Fig. 1. (a) Number of subscribers and sessions on the whole dataset. (b) CDF of number of days in which subscribers generate traffic.



(a) (b)

Fig. 2. (a) Average number of sessions per user during the week. (b) Relative Standard Deviation per parameter.

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 [5]. 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 routine behavior of people [1] and the large scale of the 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 the 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 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.

A. 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%.

As expected, there is a difference between the amount of active subscribers (a subscriber is said to be active on some day if she generates some traffic on that day) on weekdays and weekend; the highest difference is 10% which is obtained by comparing Tuesday with Saturday. On average, *the number of active subscribers is higher during the weekdays than during the weekend* (also observed in [6]). In the studied week, this average difference is 5%. In general, *the day-wise variation of the number of active subscribers is essentially decreasing as the week progresses.*

Fig. 1(b) shows the CDF of the number of active days of the subscribers within the week. 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 [6]. In addition, our analysis shows that *most subscribers generate traffic on few hours during the day.* Indeed, on 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%.

Furthermore, our analysis shows that uploaded and downloaded session volumes are similar and correlated. For instance, 35% and 38% of the sessions present upload and download volume of up to 1 MB, respectively. On the other hand, 6% and 13% of the sessions present more than 100 MB for uploaded and downloaded volume, respectively. *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.* Moreover, the Spearmans's correlation coefficient between per-session upload and download traffic is 88%. *Owing to the high correlation between the upload and download volumes, in our evaluation and traffic modeling, we take into consideration the total volume per session, i.e., the sum of the upload and download volumes during the session.*

B. Temporal Dynamics

It is common knowledge that some hours tend to be more active than others when it comes to subscribers routine 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, Fig. 2(a) show the number of sessions hourly dynamics during the week (for a complete evaluation of other parameters, refer to [7]). 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. 2(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.

Our analysis also revealed that *other parameters such as upload and download session volumes per user present the same gap between 4am to 8am and the day-wise similarity*. Interestingly, the inter-arrival time (IAT), i.e., the difference between the arrival timestamps of subsequent sessions of the same subscriber, presents the opposite behavior from 4am to 8am. It is a complementary behavior to the low average number of sessions on the same hours present in Fig. 2(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 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' 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. 2(b) 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 routine behavior. Therefore, *we select one day (namely, Wednesday) of the week to perform our extensive per-hour analysis and distinguish different profiles of users.*

III. Subscriber profiling methodology

We first define a limited number of profiles generated according to two traffic parameters: 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. Once such parameters are extracted from the sample set, a set of profiles is generated in order to describe the behavior 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.

A. 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 [8]. We then *classify* the remaining users in $\mathbb{S}$ (i.e., $\mathbb{S} - \mathbb{S}'$) into various clusters in $\mathbb{P}$. 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 $V_t^i = \sum_{k \in \tau_t^i} v_k^i$.

Similarly, the number of sessions generated by subscriber i in time slot t is $N_t^i = \sum_k \mathbb{I}(k \in \tau_t^i)$, 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., $w^{\vartheta}_{ij} = \|\vartheta^i - \vartheta^j\|$. The *number of sessions similarity* can be similarly defined as: $w^{\eta}_{ij} = \|\eta^i - \eta^j\|$.

Using the subscribers in $\mathbb{S}'$ as the vertices, and using either $w^{\vartheta}_{i,j}$ or $w^{\eta}_{i,j}$ 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^i_t and N^i_t to further classify the time slots into peak and non-peak.

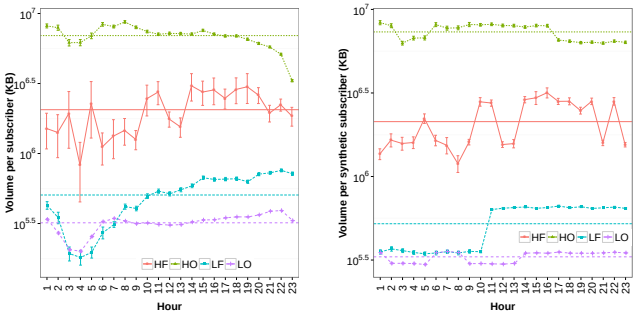
B. Subscriber Clustering and Classification

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 [9]. The hierarchical clustering algorithm we choose is the *Average Linkage clustering method*, also known as *Unweighted Pair Group Method with Arithmetic Mean (UPGMA)* [9]. 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 w^{ϑ}_{ij} or w^{η}_{ij} 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 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 [10]. Contrarily to related works that have implemented and applied very few of them [4] and in order to avoid to be biased by a specific criteria, we have implemented and used 23 stopping rules (for a complete list, refer to [7]).

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 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 w^{ϑ}_{ij} is used for the hierarchical clustering. The best number of "traffic volume"-based clusters is then determined: according to stopping criteria results, $|\mathbb{P}| = 2$ 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 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 w^{η}_{ij}. Finally, for each of these two initial clusters, two "number of sessions"-based clusters are found after the second round of stopping rules execution, totalizing four subscribers profiles. The second round ends with the classification of the remaining subscribers into the four defined profiles.



(a) (b)

Fig. 3. Volume of traffic per class per hour (a) for real and (b) for synthetic subscribers.

C. 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 behavior of a certain period of time in D. For our evaluation we choose a value of 1 hour as the time slot duration. However, the optimal size of the time slot is still an open problem [11].

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 I shows the characteristics of each of the profiles.

TABLE I
CHARACTERISTICS OF THE RESULTING PROFILES

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

In Fig. 3(a), we show the dynamics of the volume of traffic per subscribers' class per hour (for the graphics of other

traffic parameters, refer to [7]). The error bars correspond to a 95% confidence interval. For each time slot, the volume of traffic and number of sessions are calculated using V_t^i and N_t^i, respectively. For each subscriber i, the average inter-arrival time in time slot t is obtained using the following expression: $IAT_t^i = \sum_{k \in \tau_t^i}(t_{k+1}^i - t_k^i)/N_t^i$, 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. 3(a), we can see that our methodology well separates the profiles, i.e., the *occasional* and *frequent* subscribers have their values clearly separated. The uneven distribution of users per profile, e.g., LO profile contain 98% of the users may hide the importance of the other profiles. Indeed, *48% of the traffic volume is generated by the the 2% of users on the other profiles*. For each curve in this plot, 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*.

IV. 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 *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. 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: the number of sessions generated, the traffic volume associated with each of these sessions, and the inter-arrival times between the sessions (for detailed CDFs analysis, refer to [7]). 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 (Sec. III-A) 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 (Sec. III-A) and IAT_t^i (Sec. III-A), 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 [12] 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 [13] 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.

B. 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 I we see that 97.995% of the subscribers belong to LO profile, and thus with probability $q_{LO} \approx 0.97$ we assign LO profile to a synthetic user. Similarly, the probabilities of other profiles are: $q_{LF} \approx 0.001$, $q_{HO} \approx 0.018$, and $q_{HF} \approx 0.00004$. 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 (for a detailed algorithm, refer to [7]). *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, 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. For example, the number of sessions best fits to Negative binomial for all classes on peak and non-peak hours. Volume of traffic best fits to Weibull for HO on all hours and HF on peak hours, and best fits to Gamma for HF on non-peak hours, to LO and LF on all hours. IAT best fits to Gamma on all hours for HO and LO, to Log-normal on HF for all hours and LF on non-peak hours, and to Weibull on LF on peak hours. For the values of the parameters of each distribution, refer to [7].

The volume per session v_k^i (for $k \in \tau_t^i$, see Section IV) 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.

C. 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 [14]. 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, [15] 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 iff $p = p'$. We use d in order to measure the similarity between the original dataset and the synthetic dataset.

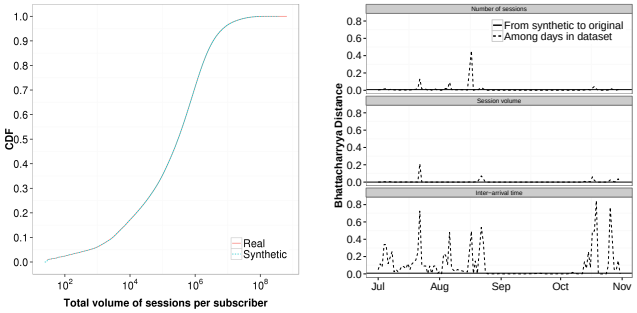
Let $\mathbb{D}$ denote a set of different time periods including D and the synthetic day denoted as D'. $\mathbb{D}$ also contains each day from 1st July to 31st October, i.e., the whole dataset. Let p_v^e denote the PDF (Probability Distribution Function) of the total volume generated by a subscriber 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. 4(a) we have depicted the CDFs corresponding to the PDFs p_ϑ^D and $p_\vartheta^{D'}$. 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 $d(p_\vartheta^D, p_\vartheta^{D'})$, the distance between the total volume distribution of the original day and the synthetic day. Then, we compute $d(p_\vartheta^D, p_\vartheta^e)$, $e \in \mathbb{D}$ but $e \neq D$, the distance between the original day and the 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. 4(b), we show the $d(p_\vartheta^D, p_\vartheta^e)$ distances. Also shown in Fig. 4(b) (horizontal solid line) is the $d(p_\vartheta^D, p_\vartheta^{D'})$ distance. Our first traffic model evaluation consists then in verifying whether the $d(p_\vartheta^D, p_\vartheta^{D'})$ is within the confidence interval of the $d(p_\vartheta^D, p_\vartheta^e)$. As can be seen in Fig. 4(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 Sec. III on the synthetic users. By doing so, we classify them and compare the per-class traffic behavior with the one created from the original dataset. Fig. 3(b) depicts the per-class behavior for the volume of traffic per session for the classified synthetic users. It is possible to see that this *result is coherent with the one for the original dataset* presented in Fig. 3(a). For instance, the behavior for peak and non-peak hours is well defined and similar to the one from the original trace.

V. 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 behavior due to their daily routines. Furthermore, we have calculated 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 investigate models to describe sessions' transfer rate and duration. Additionally, we intend to apply and evaluate our traffic generator on network planning.



(a) (b)

Fig. 4. (a) CDF of the total volume per session for real and synthetic subscribers (b) Per-parameter Bhattacharyya distances between original and synthetic trace in D, and between original trace in D and other days from the original trace.

References

[1] 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.

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

[3] 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.

[4] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," in *Proc. of IEEE Infocom*, Apr. 2014.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.