Classifying Call Profiles in Large-scale Mobile Traffic Datasets

Diala Naboulsi*, Razvan Stanica*, Marco Fiore†*
*Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA, F-69621, Villeurbanne, France – name.surname@insa-lyon.fr
†CNR – IEIIT, Corso Duca degli Abruzzi 24, 10129 Torino, Italy – marco.fiore@ieiit.cnr.it

***Abstract*—Cellular communications are undergoing significant evolutions in order to accommodate the load generated by increasingly pervasive smart mobile devices. Dynamic access network adaptation to customers' demands is one of the most promising paths taken by network operators. To that end, one must be able to process large amount of mobile traffic data and outline the network utilization in an automated manner. In this paper, we propose a framework to analyze broad sets of Call Detail Records (CDRs) so as to define categories of mobile call profiles and classify network usages accordingly. We evaluate our framework on a CDR dataset including more than 300 million calls recorded in an urban area over 5 months. We show how our approach allows to classify similar network usage profiles and to tell apart normal and outlying call behaviors.**

I. Introduction

Mobile access networks are facing an unprecedented surge in data traffic. In 2012, global mobile data traffic grew 70%, and a 13-fold rise is expected before 2017 [1]. Urban environments will pose the biggest challenge, due to the combination of ever-increasing urbanization [2] and the rise in concentration of high-end mobile devices in cities [3]. Mobile telecommunication operators are thus studying new solutions to accommodate the capacity demand in densely populated metropolitan areas. A number of approaches aiming at enhancing the access network infrastructure are under deployment, development or study. These solutions range from cognitive radio [4] to femtocell and WiFi offloading [5], from cooperative relaying [6] to re-thinking the access network architecture entirely [7].

A complementary strategy lies in improving the way network resources are utilized. There, the focus is on the characterization of network usage patterns. Indeed, individual mobile users consume network resources in significantly different ways, depending on the time at which they access the network and on the location where they do so. When aggregating user behaviors, very diverse macroscopic network utilization profiles emerge that vary over space and time. Clearly, traditional static capacity planning is hardly capable of coping with such a diversity and results in over-dimensioning and under-utilization of resources. On the contrary, accurately understanding the spatiotemporal dynamics of mobile customers' demand allows operators to allocate resources more efficiently. This is especially important in urban areas, where vast user displacements occur over short space and time scales, forcing the mobile network capacity to rapidly adapt to – or even to anticipate – demand fluctuations.

This work was supported by the French National Research Agency under grant ANR-13-INFR-0005 ABCD and by the EU FP7 ERA-NET program under grant CHIST-ERA-2012 MACACO.

Macroscopic network usage profiles are inferred from the analysis of Call Detail Records (CDRs) that describe the cell-level position and the activity of mobile network customers. CDR datasets typically include information about millions of users, collected during several months. The common practice is to process the whole data at once, and extract mobility [8], call [9], or data traffic [3] information, possibly identifying geographical or temporal patterns.

However, we argue that aggregating over whole CDR datasets has its drawbacks. First, it may outline temporal periodicities, but it does not precisely answer to the question of *which time instants show equivalent network usage profiles*. This is a critical aspect from the operator viewpoint. E.g., in the case of dynamic resource allocation, it allows determining how many different capacity configurations are needed, when each should be applied, and for how long. Second, indiscriminate aggregation of large CDR datasets may delineate global trends, but *it completely loses information on outlying user behaviors*. E.g., a Wednesday, December 8th in Lyon, France cannot be expected to yield the same mobile demand as most other Wednesdays of the year: that day, the city celebrates the *Fête des Lumieres* and its population almost doubles. Yet, in a bulk analysis of months of CDR, such an event is averaged with tens of standard Wednesdays and it disappears within the typical profile of that weekday – or, if the dataset is not large enough, it risks to introduce a bias in the results. In fact, identification of unusual network usage profiles has significant practical applications in terms of diagnosing and troubleshooting network problems or planning resource allocation in presence of specific events.

In this paper, we tackle the problem of classifying mobile demand profiles in large-scale CDR datasets. We propose an automated, parameter-free framework that allows constructing categories of call profiles from a training CDR set and classifying network usages accordingly. As a by-product of such operations, the framework can tell apart typical and outlying call behaviors. We evaluate the effectiveness of the framework on a 300-million CDR dataset including information about calls of mobile customers in the urban region of Abidjan, Ivory Coast, during a period of 5 months. We show that our framework can successfully identify call profile categories that yield meaningful social properties. Also, it is capable of correctly classifying network usages within the aforementioned categories, and of detecting unusual behaviors in the mobile demand, which we show to have clear social origins.

The paper is organized as follows. Sec. II presents previous studies focused on the analysis of CDRs. We introduce our framework in Sec. III, detailing each of its components and operation. We present in Sec. IV our reference dataset, and in Sec. V the classification results. Finally, we conclude the paper and discuss future works in Sec. VI.

II. Related work

The analysis of mobile phone data has received significant attention from researchers over the last few years. Previous works have been carried out on a wide range of subjects, including the characterization of human mobility, the analysis of network utilization, and the study of urban planning.

Human mobility studies leveraging CDR data aimed at characterizing individual and population movements [8], [10], model these movements [11] and predict them [12]. In all these works, observations are aggregated over time in order to draw conclusions on human mobility. While this approach allows avoiding the problem of sparsity in the information provided by CDRs, it also leads to mixing data referring to typical and unusual behaviors in the mobile network. Distinguishing between standard and special network activities may unveil important differences in the derived mobility patterns.

Other works focus on network utilization patterns, as we also do in this paper. Individual users' behaviors are clustered based on their calling patterns in [13], for urban planning goals. Per-user calling patterns are also the focus of [14], where the aim is however to detect users presenting anomalous call patterns. Our study differs from those above since we target a network-wide characterization rather than one focused on individual users: the problem and solutions are thus completely different. Considering works that look at the mobile network as a whole, a comparison of content consumption over a large-scale mobile network on a special event with respect to a normal day is provided in [15]. However, such analysis considers only two days, known to represent typical and unusual network usages. Our objective is to start from a large dataset comprising months of raw CDRs and infer similarities among traffic profiles so as to categorize them.

Clearly, categorizing mobile traffic profiles is not a trivial task itself. First, a definition of similarity is required in order to place network usages in the same group. Previous research efforts have mainly focused on the total traffic volume when characterizing users' behaviors [9]. Studying this metric only reflects large positive or negative variations in the total mobile traffic volume over the studied region, and does not account for precise geographical variations within it. Other works consider a higher spatial granularity, by studying aggregated traffic volumes over a group of base stations [16], [17], or by looking at each base station independently [18]. However, this still does not provide any information regarding the distribution of the mobile traffic volume among different areas in the region of interest. In fact, understanding how the volume is distributed over different areas of a city is compulsory for the study of network utilization patterns. The works in [19] and [20] captured that aspect by considering the traffic volume in each area of a specific region to be normalized with respect to the total traffic volume in the region. However, these studies are limited to the presentation of the normalized volume, and do not consider any measure of similarity between traffic patterns nor provide a classification of network usage profiles.

This work is the first to introduce similarity measures for the comparison of traffic patterns in terms of volume variations and volume distributions. Moreover, we integrate this original approach into a complete classification framework capable of telling apart typical and outlying mobile traffic profiles.

III. Framework

In this section, we present our framework for the classification of mobile network usage profiles. The framework runs on *snapshots* of the mobile demand extracted from raw CDRs. As the name suggests, a snapshot is a representation of the load generated by mobile users on the access network at a given time instant. Apart that, we do not impose any constraint on the way snapshots are defined: they can describe the traffic volume at every second or averaged over longer time intervals, at each base station or aggregated over larger geographical areas, and for one or multiple types of services (e.g., voice calls, short text messages, Internet-based applications, etc.). In the following, we will denote as $\mathbb{T}$ the set of time intervals snapshots refer to: values in this set will be thus used to index the snapshots. Similarly, $\mathbb{Z}$ will indicate the set of geographical areas over which traffic volumes are aggregated[1]. The choice of $\mathbb{T}$ and $\mathbb{Z}$ may depend on the level of detail of the available CDRs or on the target of the study, yet our framework is general enough to accommodate it. We will provide a practical example of snapshot definition when introducing the dataset employed for our performance evaluation, in Sec. IV.

Once snapshots are defined and extracted from the CDR dataset, the framework processes them through four phases. The first three phases aim at defining a limited number of network usage categories by analyzing a training set of snapshots, and their workflow is depicted in Fig. 1. The fourth phase allows to classify additional usage profiles into the categories above. The different phases are detailed in the remainder of this section.

A. Snapshot graph

In the first step, a subset $\mathbb{T}' \subseteq \mathbb{T}$ of snapshots is selected as the training set over which the categories of network usage profiles are defined. The choice of $\mathbb{T}'$ mainly depends on the available CDR dataset. As an example, an operator may choose to use snapshots retrieved from the past one-year history to train the framework, so as to be able to classify the following network usage profiles as they are recorded[2].

Snapshots in $\mathbb{T}'$ are then mapped to the vertices of an undirected weighted graph $G(\mathbb{T}', \mathbb{E})$ that we dub *snapshot graph*. In the definition above, $\mathbb{E} = \{e_{ij} \mid i, j \in \mathbb{T}',\ i \neq j\}$ is the set of edges e_{ij} between any two snapshots i and j of the training set $\mathbb{T}'$. Therefore, the snapshot graph is a clique, as all pairs of vertices share an edge. Each edge e_{ij} is assigned a weight w_{ij}, which is a measure of the similarity between the network usage profiles in snapshots i and j. The way such similarity is measured plays an important role in the framework operation. We propose two different definitions of usage profile similarity that capture complementary facets of mobile traffic dynamics. They are detailed next.

Traffic volume similarity Given a snapshot $i \in \mathbb{T}'$, we use v_i^z to indicate the mobile traffic volume[3] observed in the geographical area $z \in \mathbb{Z}$.

[1] At the highest spatial granularity level, $\mathbb{Z}$ maps to the set of base stations.

[2] Having a limited CDR dataset, we rather test multiple training sets: we show that the framework yields consistent results for different $\mathbb{T}'$'s in Sec. V.

[3] As previously stated, our definition of mobile traffic volume is as general as possible. Depending on the available CDR dataset(s) and on the target of the study, one can consider overall, inbound or outbound traffic, as well as traffic generated by all or just some specific applications.

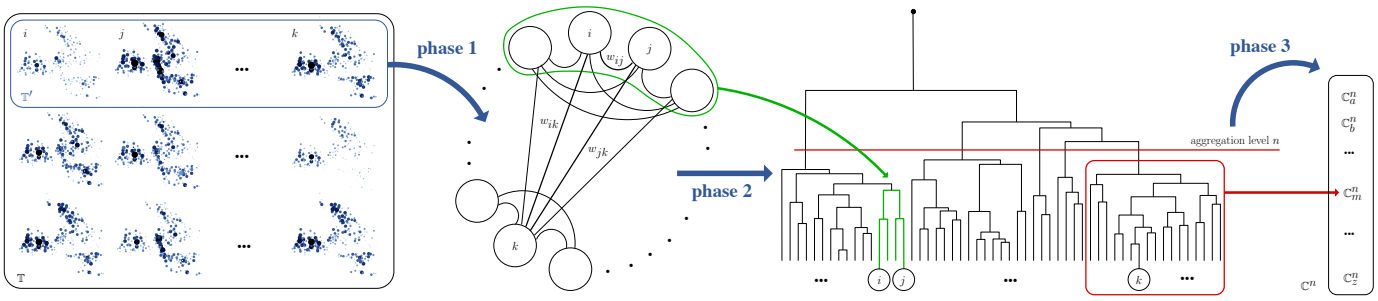


Fig. 1. Workflow of the framework for the definition of categories of network usage profiles. Phase 1: construction of the snapshot graph from snapshots (portrayed here as geographical plots of the mobile traffic volume) in the training set $\mathbb{T}'$. Phase 2: iterative aggregation of graph vertices into a dendrogram structure. Phase 3: identification of the clustering level n granting the maximum separation between the groups of snapshots. The resulting clusters $\mathbb{C}^n_m \in \mathbb{C}^n$ are mapped to network usage profile categories.

The easiest way to compare the traffic volume recorded in two snapshots i and j is to look at the difference of the overall amount of exchanged data, i.e., $\sum_{z\in\mathbb{Z}} v_i^z - \sum_{z\in\mathbb{Z}} v_j^z$, or at measures directly derived from it. In fact, this is a very common approach in the literature.

However, while it permits to identify large positive or negative variations in mobile traffic, this metric does not account for spatial diversities. Thus, we introduce a *traffic volume similarity* measure $\mathcal{V}$ that accounts for traffic volume variations between two snapshots i and j by distinguishing among the different geographical areas. If the traffic volume similarity is used to determine the weights of the snapshot graph edges, then:

$$w_{ij} = \frac{1}{\sqrt{\sum_{z\in\mathbb{Z}} (v_i^z - v_j^z)^2}}.$$

If we consider that we have only one area in $\mathbb{Z}$, mapping to the whole region under study, then $\mathcal{V}$ maps to the total volume variation above. On the other hand, if we divide the region of interest into a significant number of areas, $\mathcal{V}$ can capture the spatial diversity in the mobile traffic.

Traffic distribution similarity. The $\mathcal{V}$ metric alone does not provide a complete description of the calling profile. While it accounts for absolute variations of mobile traffic over separate areas, it overlooks how the traffic is distributed among such areas. We thus introduce a second measure $\mathcal{D}$, named *traffic distribution similarity*, that is capable of showing how the mobile traffic is divided among different areas. The weight between two snapshots i and j is then obtained as:

$$w_{ij} = \frac{1}{\sqrt{\sum_{z\in\mathbb{Z}} \left(v_i^z/V_i - v_j^z/V_j\right)^2}}\ ,\ V_i = \sum_{z\in\mathbb{Z}} v_i^z \quad \forall i \in \mathbb{T}'.$$

There, V_i represents the total traffic volume recorded in the whole studied region during snapshot i. Thus, $\mathcal{D}$ considers the normalized volume at each area $z \in \mathbb{Z}$, rather than the absolute one as done in the case of $\mathcal{V}$. This allows to capture how the traffic is distributed over the region, independently of its absolute volume.

In our evaluation, we will use both $\mathcal{V}$ and $\mathcal{D}$ as snapshot similarity measures, as they play complementary roles in the identification of network usage profiles. This implies that one snapshot graph will be built for each measure, and that the next phases will be performed separately on the two graphs.

B. Snapshot aggregation

The snapshot graph is used in the second phase as a base for the definition of a list of potential categories of network usage profiles. Thus, the goal here is to identify all possible partitionings of the graph vertices, i.e., snapshots, that display similar mobile traffic conditions. To that end, a hierarchical clustering algorithm iteratively aggregates graph vertices in the snapshot graph into larger clusters, and organizes them into a dendrogram structure, as in Fig. 1.

We selected the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [21] – also known as mean or average linkage clustering – as the hierarchical clustering method. UPGMA relies on an agglomerative clustering approach that starts from singleton clusters including one graph vertex each. Then, at every iteration, the algorithm merges the two clusters that are deemed to share the strongest ties: from our viewpoint, this means to aggregate the groups of snapshots that show the highest level of similarity in terms of network usage profiles. The procedure is carried out until a single cluster partition matching the whole graph is obtained.

Specifically, at iteration n of UPGMA (i.e., at the aggregation level n of the resulting dendrogram), graph vertices are placed into disjoint clusters $\mathbb{C}^n_k \subseteq \mathbb{T}'$ forming a set $\mathbb{C}^n = \{\mathbb{C}^n_k\}$. The algorithm computes the average distance between each pair of clusters $\mathbb{C}^n_k$ and $\mathbb{C}^n_h$ in $\mathbb{C}^n$ as:

$$d^n_{kh} = \frac{1}{|\mathbb{C}^n_k| \cdot |\mathbb{C}^n_h|} \sum_{i\in\mathbb{C}^n_k, j\in\mathbb{C}^n_h} \frac{1}{w_{ij}}.$$

Once such a value has been computed for all pairs, the two clusters $\mathbb{C}^n_k$ and $\mathbb{C}^n_h$ having the smallest average distance are joined into a new cluster $\mathbb{C}^{n+1}_m$, and the new set $\mathbb{C}^{n+1}$ is defined accordingly. As a result, snapshots are organized in a dendrogram structure outlining all the partitionings that progressively gather similar network usage profiles.

C. Network usage profile categories

In the third phase, we study the dendrogram generated by UPGMA and determine the clustering level yielding the best separation among the groups of snapshots. The resulting clusters will become our network usage profile categories.

To that end, we consider several indices, also known as stopping rules for clustering algorithms. These indices are calculated at each level of the dendrogram to quantify the

separation among clusters. We referred to the extensive survey in [22] to make our choice of stopping rules. Specifically, we implemented and tested four different top-ranking indices among the 30 candidates proposed in the literature and compared in the aforementioned study: they are Calinski and Harabasz, Beale, Duda and Hart, and the C indices. Due to space constraints, in this paper we only introduce and present results for the first two stopping rules above. We stress however that we obtained consistent outcomes with all indices, as they converged to the same clustering of graph vertices.

Calinski and Harabasz index. The Calinski and Harabasz (CH in the remainder of the paper) index is calculated for a generic level n of the dendrogram as follows:

$$CH^n = \frac{B^n}{P^n} \cdot \frac{|\mathcal{T}'| - |\mathbb{C}^n|}{|\mathbb{C}^n| - 1}\ , \text{ with } B^n = \sum_{\mathbb{C}_k^n \in \mathbb{C}^n} |\mathbb{C}_k^n| \left(\frac{1}{w_{\overline{c}_k^n, \overline{s}}} \right)^2 ,$$

$$\text{and } P^n = \sum_{\mathbb{C}_k^n \in \mathbb{C}^n} \sum_{i \in \mathbb{C}_k^n} \left(\frac{1}{w_{i, \overline{c}_k^n}} \right)^2 .$$

There, $\overline{c}_k^n$ is the *center* of cluster $\mathbb{C}_k^n$: it is a synthetic snapshot representing the center of mass of the cluster, obtained by averaging the traffic volume recorded over all the snapshots of the cluster. Similarly $\overline{s}$ is a synthetic snapshot representing the center of all snapshots in the training set $\mathbb{T}' = \bigcup_{\mathbb{C}_k^n \in \mathbb{C}^n} \mathbb{C}_k^n$.

Element B^n is a measure of how separate clusters in $\mathbb{C}^n$ are, as it sums up the distances between the center of each cluster and the center of all the training set data. Conversely, P^n evaluates the proximity of snapshots belonging to the same cluster, by leveraging the distance between every snapshot i in a cluster $\mathbb{C}_k^n$ and the center of the cluster $\overline{c}_k^n$. Clearly, B^n and P^n respectively decrease and increase as n grows: the second factor compensates this, as it becomes larger as the number of clusters $|\mathbb{C}^n|$ is reduced, and allows for a fair comparison between different aggregation levels.

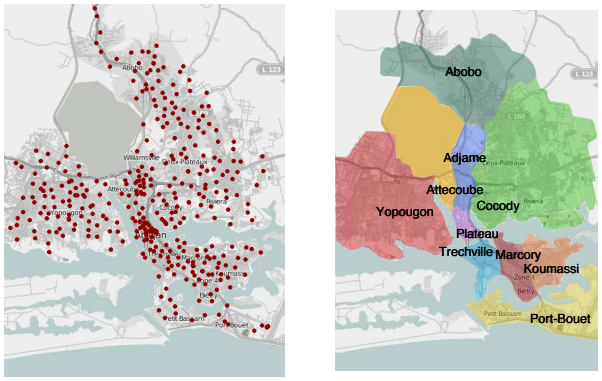
Overall, the CH index compares the distance among clusters B^n to the level of internal cohesion of clusters P^n to determine the quality of clustering: the higher the value of the index, the better the clustering. Therefore, the dendrogram level n retaining the highest CH index value is the one that grants the best separation among clusters.

Beale index. The Beale index represents the F-ratio of a statistical F-test that accepts or rejects the merging of two clusters at level n into a new cluster at level $n+1$. Suppose that level-n clusters $\mathbb{C}_k^n$ and $\mathbb{C}_h^n$ merge to form a cluster $\mathbb{C}_m^{n+1}$ at level $n+1$. Then, the Beale index would be:

$$F^n = \frac{P_m^{n+1} - (P_k^n + P_l^n)}{(P_k^n + P_l^n)} \Bigg/ \left(\frac{|\mathbb{C}_m^{n+1}| - 1}{|\mathbb{C}_m^{n+1}| - 2} \cdot 2^{2/|\mathbb{Z}|} - 1 \right)$$

$$\text{with } P_k^n = \sum_{i \in \mathbb{C}_k^n} \left(\frac{1}{w_{i, \overline{c}_k^n}} \right)^2 .$$

This F-ratio considers the variation of distance among snapshots within the two original cluster and that among the same snapshots when they are grouped within the same cluster. F^n is compared to the critical value F_{crit}^n returned by an F-distribution $F\left(|\mathbb{Z}|, (|\mathbb{C}_m^{n+1}| - 2)|\mathbb{Z}|\right)$ at a significance level of 5%. The null hypothesis that a the clustering quality at level $n+1$ is better than that at level n is rejected if $F^n > F_{crit}^n$. Therefore, the dendrogram level n corresponding to the best clustering quality is that for which $F^n - F_{crit}^n$ is maximum.



(a) (b)

Fig. 2. (a): Base station deployment in Abidjan. (b): Communes of Abidjan.

D. Snapshot classification

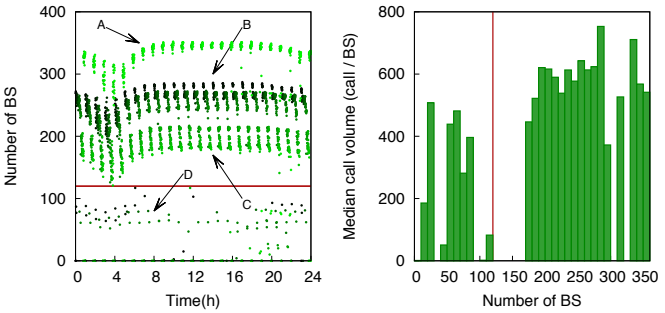
Stopping rules allow us to define the aggregation level at which clusters of snapshots show the best tradeoff between intra-cluster cohesion and inter-cluster separation. We thus retain the corresponding clustering for our definition of network usage profile categories, as portrayed in Fig. 1.

Once the set of categories is identified over snapshots in $\mathbb{T}'$, we can classify the remaining snapshots in $\mathbb{T} \setminus \mathbb{T}'$ accordingly. To that end, we assign each unclassified snapshot to a category, via the k-means algorithm [21]. While k-means is commonly known as a partitional clustering algorithm, here we use it as a classification technique. As a clustering algorithm, k-means assigns a set of objects to a predetermined number k of clusters. To that end, it starts from an arbitrary initial assignment of all objects to the k clusters.

In our case, instead, we consider the categories defined at the end of the third phase above as the initial partitioning of data for k-means, and apply the algorithm to snapshots in $\mathbb{T} \setminus \mathbb{T}'$. At each iteration, k-means allows then to choose the best category for the current snapshot. The similarity measure used by k-means is the average distance between the yet unclassified snapshot i and each cluster $\mathbb{C}_k^n \in \mathbb{C}^n$, i.e., $1/w_{i, \overline{c}_k^n}$. The algorithm then assigns the snapshot to the category for which such measure is the smallest.

IV. Dataset

We test our framework on a dataset provided by Orange within the context of the D4D Challenge [23]. The dataset is based on anonymized CDRs of 5 million Orange customers in Ivory Coast, and it presents the mobile traffic volume in terms of number of voice calls exchanged between any two Orange base stations in the country, aggregated for each hour of the observation period. The information covers over 5 months, from December 5th, 2011 to April 22nd, 2012. As our interest is on urban environments, we focus on the city of Abidjan, the economic capital of Ivory Coast and a highly populated 500-km^2 area with more than 4 million inhabitants. We filter the dataset by keeping only the information involving the antennas



(a) Number of base stations (b) Median call traffic

Fig. 3. (a): Number of active antennas in each snapshot of the 5-month dataset, for the different hours of the day. (b): Median volume per base station as a function of the number of active base stations, aggregated over the 5-month period for hours between 10:00 and 20:00.

in Abidjan, which leaves us with 364 base stations whose deployment is shown in Fig. 2(a).

As the D4D dataset features a one-hour granularity, we consider snapshots to include aggregate information over each hour. Thus, our set $\mathbb{T}$ contains over 3600 snapshots, each describing the network usage during a specific hour. The set of geographical ares $\mathbb{Z}$ over which traffic volumes are aggregated is mapped to the set of *communes* of the city, shown in Fig. 2(b). While we acknowledge that different options might be envisioned, this spatial aggregation is intuitive – as it is based on topological criteria – and proved to be effective in terms of results obtained. Also, the available dataset only contains voice call volumes, and therefore all our results refer to voice traffic.

Before moving to the discussion of the results, an important remark is in order. During our analysis, we noted that the information regarding the 364 Abidjan antennas is not always present for the entire observation period. In Fig. 3(a), we check the evolution of the number of antennas over the 5 months. Each point on this figure represents a specific snapshot extracted from the dataset: the x coordinate is the day hour the snapshot maps to, while the y axis corresponds to the number of active antennas for which some traffic is recorded during the snapshot. We distinguish between the different days with a color contrast degradation, such that dark green maps to the first day in the dataset and light green maps to the last one.

Four different situations can be detected in Fig. 3(a). The first one, labeled as A, includes all the 364 antennas and covers the period between March 28th, 2012 to April 22nd, 2012. The second period, B, goes from December 7th, 2011 until February 21st, 2012, with around 250 antennas. Period C, featuring around 170 stations, spans between February 22nd, 2012 and March 27th, 2012. Finally, a fourth scenario, tagged as D, can be observed on the figure, with a very small number of antennas. Unlike the three previous cases, D does not map to a specific time interval, but includes relatively short intervals coming from the entire 5-month period. As snapshots in D are uniformly distributed over the 24 hours, they cannot be explained by the normal reduction in cellular traffic at night[4]. After investigating, we found out that the first three behaviors can be explained by different collection periods during which voice call traffic was recorded for intersecting yet different subsets of base stations. On the contrary, the behavior labeled as D is the consequence of technical problems encountered by the operator and occasional electricity failures in Abidjan.

Such inconsistencies within the dataset led us to ask ourself which portion of the snapshots actually provides reliable information on the mobile traffic. Fig. 3(b) shows the median call traffic volume per base station versus the number of active base stations, obtained from all the snapshots in the 5-month dataset falling between 10:00 and 20:00 (in order to eliminate any bias introduced by low-traffic night hours). From the figure, we observe that the three periods A, B and C map to a consistently high median volume per base station, while period D presents highly variable median values with a lower average. Therefore, the different number of active base stations does not seem to influence the voice call activity in periods A, B and C. Conversely, after carefully studying specific dates within period D, we concluded that the irregular traffic volumes in snapshots belonging to D are not representative of actual network usage profiles. Indeed, these situations are produced by major technical problems in the network and do not reflect the behavior of mobile users. Therefore, we eliminate from $\mathbb{T}$ and do not consider in our analysis all the snapshots that fall in the D category, i.e., for which less than 120 active antennas are recorded.

It is noteworthy that our decision of analyzing snapshots from the A, B and C periods together implies that two different snapshots i and $j \in \mathbb{T}$ can contain a different number of base stations. In order to fairly compute the similarity of snapshots with different number of base stations, we only consider base stations that appear in both snapshots in the calculation of w_{ij}.

V. Results

We applied the framework introduced in Sec. III to the dataset detailed in Sec. IV. Here, we present the outcome of the evaluation. First, we discuss how our results stress the complementarity of the two similarity measures $\mathcal{V}$ and $\mathcal{D}$, in Sec. V-A. Then, we present the categories resulting from the first three steps of the framework, in Sec. V-B, and the corresponding classification, in Sec. V-C. Finally, we report on the outlying behaviors identified in the process, in Sec. V-D.

A. Complementarity of $\mathcal{V}$ and $\mathcal{D}$

The first question we want to answer is whether the intuition behind the definition of two different similarity measures is correct. To investigate this aspect, we focus on the relationship between the similarity values w_{ij} returned by $\mathcal{V}$ and $\mathcal{D}$ for each pair of snapshots $i, j \in \mathbb{T}$. To this end, we calculate the Pearson product-moment correlation coefficient (PPMCC) [24] to measure the linear correlation between the two measures. The value of the PPMCC lies in the interval [-1,1], with a value of -1 indicating a negative linear correlation between variables, and a value of 1 implying a perfect linear correlation between data points. A value close to 0 indicates that there is no linear correlation between the considered variables.

In our case, the calculation of PPMCC led to a value of 0.239, which means that no actual correlation exists between $\mathcal{V}$ and $\mathcal{D}$. This can be also observed in the scatterplot in Fig. 4,

[4]Natural call volume reductions can be observed for periods A, B and C between 3:00 and 6:00, when very few individuals use the cellular network.

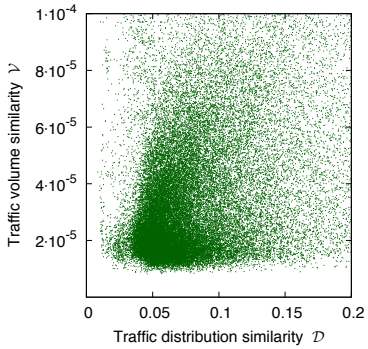
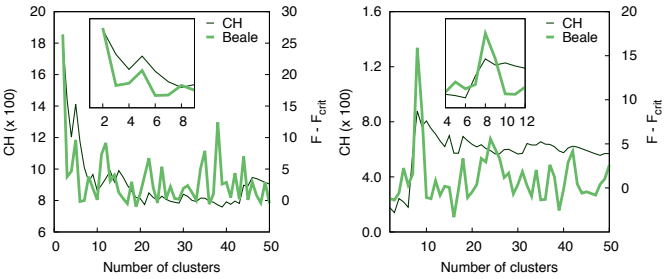


Fig. 4. Scatterplot of measures $\mathcal{V}$ and $\mathcal{D}$ for each pair of snapshots $i, j \in \mathbb{T}$.



(a) Similarity measure $\mathcal{V}$ (b) Similarity measure $\mathcal{D}$

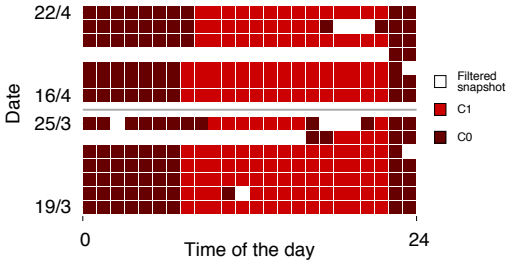
Fig. 5. CH and Beale clustering indices versus the number of clusters (i.e., the dendrogram aggregation level) for a training set of two weeks. (a): Dendrogram built using the traffic volume similarity measure $\mathcal{V}$. (b): Dendrogram built using the traffic distribution similarity measure $\mathcal{D}$.

where each dot represents a snapshot pair, and the x and y coordinates match the associated values of $\mathcal{V}$ and $\mathcal{D}$. The low correlation of the similarity measures lets us advocate in favor of the importance of considering them both when characterizing large-scale mobile traffic behaviors. Relying only on the total volume of a snapshot (or on measures derived from it) to compare mobile traffic profiles is a common practice nowadays, however we argue that is not sufficient for all situations. The results we present next confirm this.

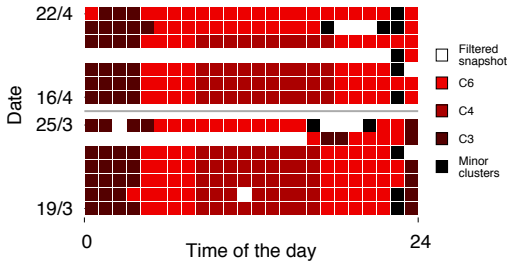
B. Call profile categories

As explained in Sec. III, our framework requires a training set $\mathbb{T}' \subseteq \mathbb{T}$ from which to derive the different categories of network usage profiles. We decided to train the framework on a two-week period: on the one hand, this allows to cover the strong weekly periodicity of human activities; on the other, considering two weeks instead of one is a safety measure in case one of the two weeks presented special events that can bias the process. We thus selected different random pairs of weeks from the 5-month dataset, and run the first three phases of the framework on each such pair. Interestingly, the number of obtained categories and their content yield negligible differences under all pairs of weeks used as training set. This proves the robustness of the framework to choice of training set, and validates our decision to consider two weeks as the training set duration.

Samples of the category selection process are shown in Fig. 5. The plots show the evolution of the two clustering indices introduced in Sec. III-C, the CH and Beale indices, versus the number of clusters. The latter is an expression



(a) Traffic volume similarity measure $\mathcal{V}$



(b) Traffic distribution similarity measure $\mathcal{D}$

Fig. 6. Content of mobile traffic profile categories defined on the training set $\mathbb{T}'$ composed of the two weeks March 19th–25th and April 16th–22nd, 2012. Each square represents one snapshot, whose category maps to a color. Empty squares are snapshots that were filtered out from the dataset as from Sec. IV.

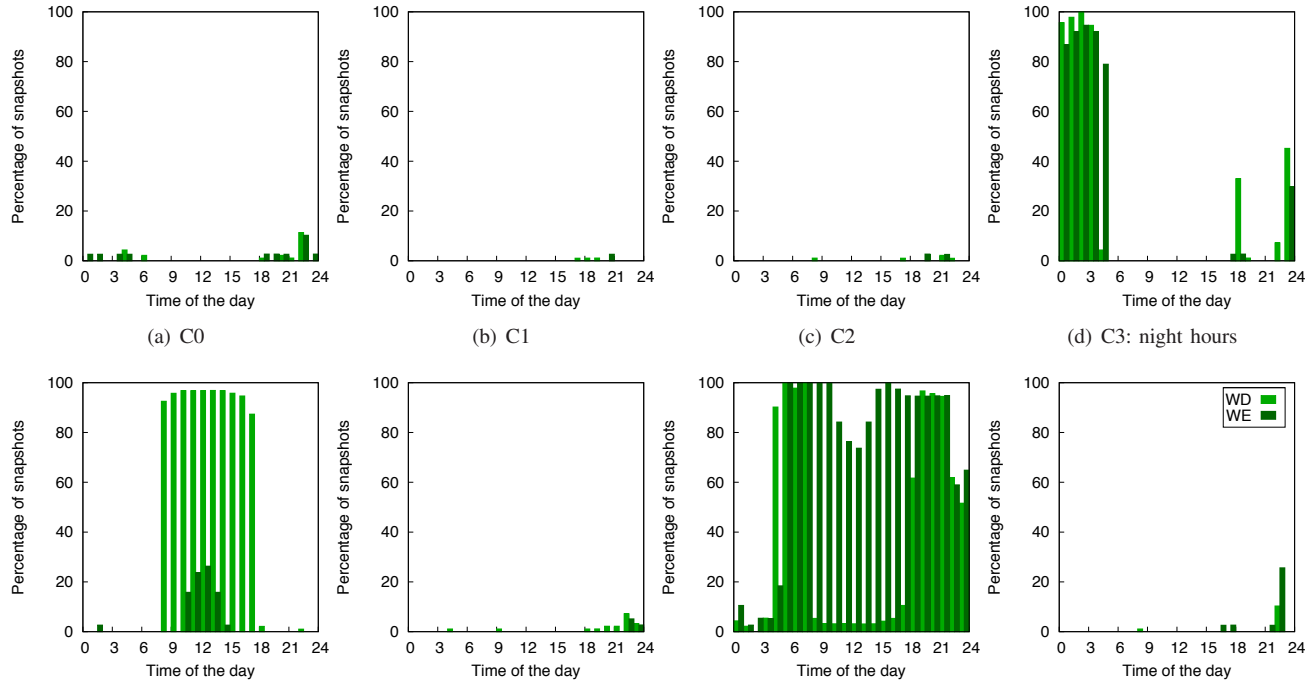
of the aggregation level, as aggregating clusters is equivalent to reducing their number. Fig. 5(a) refers to the dendrogram obtained when using the traffic volume similarity measure $\mathcal{V}$, and Fig. 5(b) to that obtained when using $\mathcal{D}$. The indices agree that the best separation between snapshots in $\mathbb{T}'$ is obtained at *two* clusters for $\mathcal{V}$ and at *eight* clusters for $\mathcal{D}$, as CH^n and $F^n - F^n_{crit}$ reach their maxima at these values.

We present the structure of the categories found over a sample training dataset for $\mathcal{V}$ and $\mathcal{D}$ in Fig. 6(a) and Fig. 6(b), respectively. We note that the two categories identified based on $\mathcal{V}$ clearly separate times with a lower activity, i.e., hours between 22:00 and 7:00, and times with a higher traffic, i.e. hours between 8:00 and 21:00.

More interestingly, for the case of the $\mathcal{D}$ metric, we can observe that the snapshots of the training set belong to three major clusters out of the eight identified. The first category includes the snapshots of the night hours, between 23:00 and 4:00, characterized by a low traffic generated in the residential areas of the city. The second category includes daytime snapshots from the weekdays, i.e., hours between 10:00 and 17:00, Monday to Friday: these snapshots show higher mobile traffic in the office and university areas. The third major category contains snapshots from weekend days, i.e., Saturdays and Sundays, as well as from early morning (5:00–9:00) and evening (19:00–22:00) hours of weekdays: the corresponding network usage is that of a high traffic volume generated in the residential areas. Also, five minor clusters appear, including a very small number of snapshots each.

C. Call profile classification

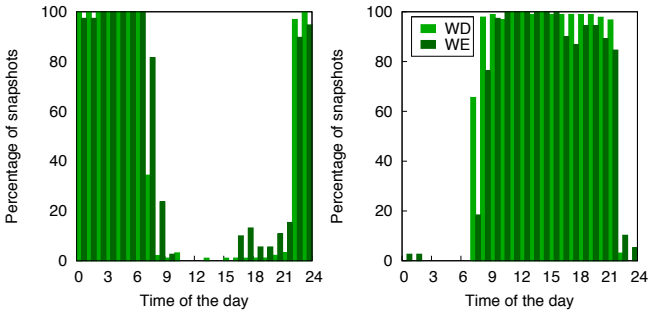
Once categories have been defined on snapshots of a two-week training set $\mathbb{T}'$, all of the remaining snapshots in $\mathbb{T} \setminus \mathbb{T}'$ (corresponding to 20 weeks of snapshots in our dataset)



(a) C0 (b) C1 (c) C2 (d) C3: night hours

(e) C4: weekdays day hours (f) C5 (g) C6: weekends, early/late weekdays (h) C7

Fig. 8. Classification of the 5-month data using the similarity measure $\mathcal{D}$. The three major categories C3, C4 and C6 are tagged with matching typical hours.



(a) Category one: low activity (b) Category two: high activity

Fig. 7. Classification of the 5-month data using the similarity measure $\mathcal{V}$.

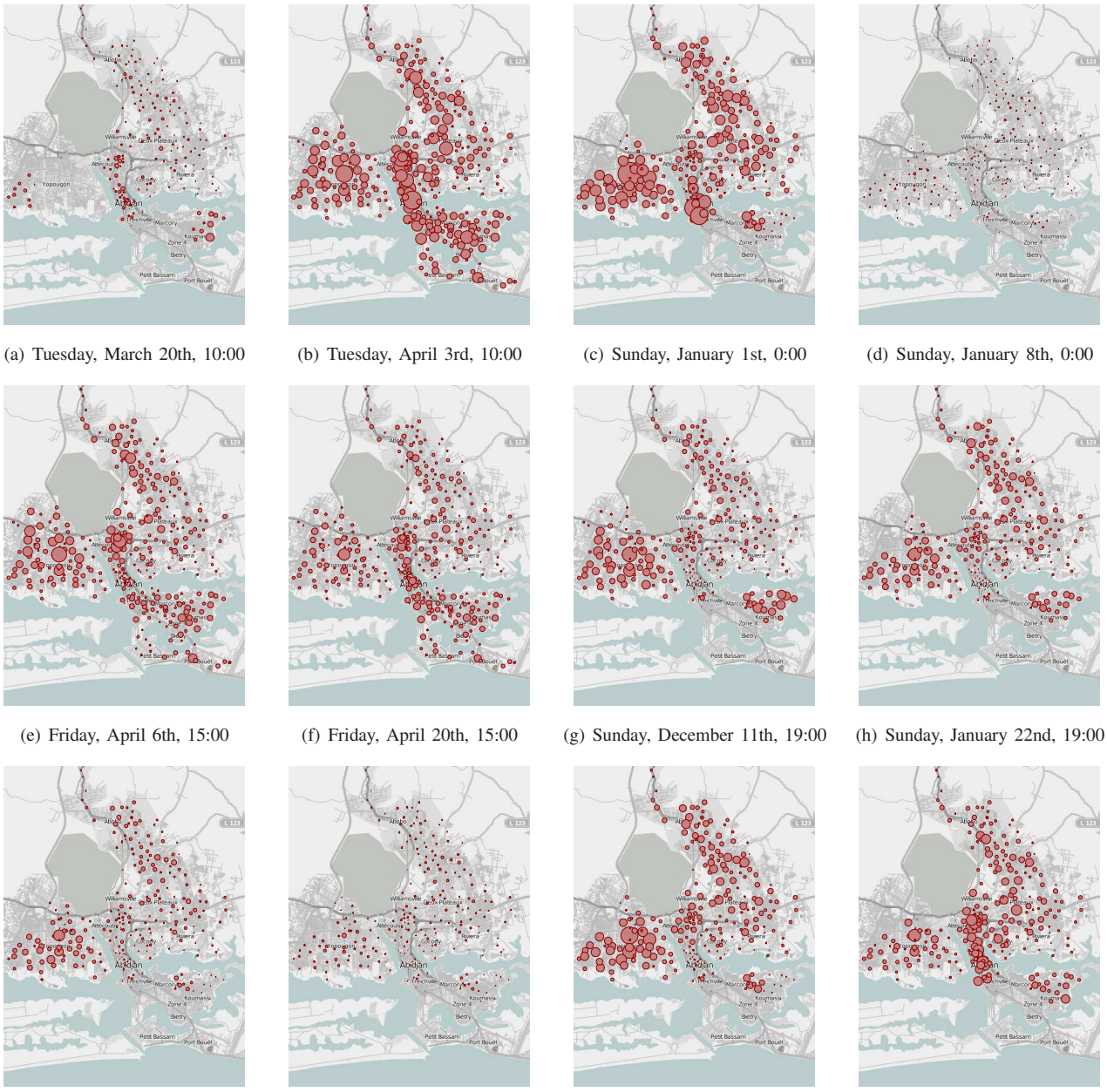
are classified accordingly. Here, we look at the content of the different categories once such a classification process is completed. We plot in Fig. 7 and Fig. 8 the content of the categories defined under the measures $\mathcal{V}$ and $\mathcal{D}$, respectively. Since we observed that categories are strongly related to the day hour and to day type (i.e., week or weekend days), we portray the results so as to facilitate the visualization of such a structure. Therefore, each plot represents one category: it reports the percentage of snapshots in $\mathbb{T}$ corresponding to a certain time of the day and to either a week or weekend day, which fall into that category.

We can observe in both Fig. 7 and Fig. 8 that the categories retain their initial structure, as most snapshots in the 5-month dataset are classified in what can be considered their typical category. However, we notice that some snapshots join categories that differ from those they supposedly belong to, i.e., yield unusual network usage profiles. These snapshots thus represent outlying behaviors, which we discuss in detail next.

D. Call profile outliers

Focusing on the the two categories obtained with the similarity measure $\mathcal{V}$, we observe that some snapshots at day time hours, such as 10:00 and 16:00, join unexpectedly the low-activity category in Fig. 7(a). Clearly, these snapshots present outlying behaviors, in terms of number of calls. Some of these outliers can be either explained by minor technical problems in the network or secondary electricity failures – and this despite our efforts in filtering such snapshots described in Sec. IV. Indeed, in some cases, these external problems happen even when more than 120 base stations are present in the dataset, and obviously affect the call volume. As an example, we consider the case of Tuesday, March 20th at 10:00, whose call volume is portrayed in Fig. 9(a). There, each dot maps to one base station in the snapshot, and the dot size is proportional to the volume of calls managed by the base station. Comparing this snapshot with another one classified as a typical Tuesday at the same time (e.g. Tuesday, April 3rd at 10:00, in Fig. 9(b)), we can notice that an important number of antennas is missing from the dataset on March 20th. We consider such a result as a demonstration that an automated framework can identify unusual network behaviors at a much finer grain than an aggregate data analysis.

Concerning the outliers falling in the second cluster in Fig. 7(b), i.e., low-activity hours showing an uncommon surge in mobile traffic, these are mostly related to special events. This is the case, e.g., of New Year's Eve, portrayed in Fig. 9(c), when people are making calling much more than on a typical Sunday at the same time (e.g. January 8th, in Fig. 9(d)). Based on the same principle, we could detect outliers, e.g., on Christmas' Eve, on the day of the quarter final and on that of the final football games of the Africa Cup of Nations.



(a) Tuesday, March 20th, 10:00 (b) Tuesday, April 3rd, 10:00 (c) Sunday, January 1st, 0:00 (d) Sunday, January 8th, 0:00

(e) Friday, April 6th, 15:00 (f) Friday, April 20th, 15:00 (g) Sunday, December 11th, 19:00 (h) Sunday, January 22nd, 19:00

(i) Saturday, December 24th, 23:00 (j) Saturday, December 17th, 23:00 (k) Sunday, December 25th, 19:00 (l) Monday, February 6th, 09:00

Fig. 9. Call volumes in Abidjan for different snapshots. In each plot, one base station maps to a dot, whose size is proportional to the voice traffic volume.

As far as the classification based on the $\mathcal{D}$ measure is concerned, in Fig. 8 we can notice multiple situations where a snapshots diverge from the expected behavior.

At times, snapshots deemed to end in one of the three major categories C3, C4 and C6 fall into another major category. Such is the case, for example, of Friday, April 6th at 15:00, depicted in Fig. 9(e), with respect to a typical behavior portrayed in Fig. 9(f), representing Friday, April 20th at 15:00. This outlying behavior happens to be the Good Friday, whose afternoon is a public holiday. This explains the fact that the snapshot is classified together with weekend snapshots in C6: it shows an increase in call volume in residential areas (Yopougon, Adjame, and Abobo), and a volume decrease in Plateau, the largest office and commercial area of the city. The same is true of the other outliers falling in C6: e.g., the whole day of Easter on Monday, April 9th, is classified under the weekend profile by the framework.

Similarly, outliers falling in the weekday daytime category C4 are related to special events involving calling activities during the weekend that are close to those observed on normal weekdays over residential and working regions. Finally, outliers joining the night hours category C3 reflect a reduced level of calls in the network.

Other snapshots diverge from the typical behavior by joining minor clusters. These snapshots are related to special events that do not concern the whole population of Abidjan, but are localized. Thus, these event imply mobile traffic distributions that differ from what can be observed in any *typical* major

TABLE I
LIST OF OUTLYING SNAPSHOTS, ACCORDING TO THE CLASSIFICATION PROVIDED BY THE MEASURE $\mathcal{D}$

Date	Category	Expected category	Event
Sunday, January 1st, 0:00	C6	C3	New Year's Eve
Saturday, February 4th, 13:00	C4	C6	The Birth of the Prophet
Monday, April 9th, 10:00 – 17:00	C6	C4	Easter Monday
Friday, April 6th, 15:00 – 17:00	C6	C4	Good Friday
Wednesday, December 7th, 18:00	C4	C6	Anniversary of the death of Felix Houphouet Boigny
Saturday, January 7th, 11:00	C4	C6	Hilary Clinton and Kofi Annans visit to Abidjan
Tuesday, March 13th, 18:00	C5	C6	Election of National Assembly President and Prime Minister
Sunday, December 11th, 19:00	C0	C6	New parliament election
Sunday, February 12th, 23:00	C3	C6	Africa Cup of Nations final

category. As an example, the snapshot in Fig. 9(g) falls in a minor category and happens to occur on the day of the election of a new parliament, i.e., Sunday, December 11th at 19:00. Comparing it with the typical behavior of the same day of the week at a different date (e.g., Sunday, January 22nd at 19:00, in Fig. 9(h)), we notice that it is detected due to the increase of the calling volume in the regions of Yopougon and Koumassi, i.e., the residential areas where most people live.

In Table I, we present a list of the outliers detected by the classification obtained with measure $\mathcal{D}$, that we were able to relate to special events. The table also reports the category where each outlier was detected, the one where it was expected to end, and the social reason behind the outlying behavior.

Another important result concerning $\mathcal{D}$ is the fact that snapshots with a different call volume, but with a similar traffic distribution, were assigned to the same category. For example, the snapshots presented in Fig. 9(i) and Fig. 9(j) both belong to the C3, showing that, although people are making more calls on Saturday, December 24th at 23:00, the behavior is uniform over the entire city, and these calls come from places that are usual for a typical week-end evening. In fact, these two snapshots were placed in different clusters based on the $\mathcal{V}$ metric: in that case, December 24th at 23:00 was considered an outlying behavior due to the increased traffic volume.

We also detected cases where snapshots were belonging to the same category based on $\mathcal{V}$ but were classified in different categories based on $\mathcal{D}$. This means that, for similar levels of volume, one can observe several volume distributions. Such is the case of the snapshots appearing in Fig. 9(k) and Fig. 9(l).

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework to characterize mobile traffic profiles, allowing to build categories of network usages. We evaluated our framework on a large-scale dataset of voice calls in Abidjan, Ivory Coast. The results show how the classification produced by the framework consistently clusters calling profiles with common features that depend on the similarity measure employed. As an interesting by-product, the framework successfully identifies unexpected traffic profiles.

In our future work, we aim at increasing the granularity of the analysis by considering the internal structure of the categories. We also consider extending our study by accounting for additional measures that go beyond traditional traffic volumes and by evaluating the framework with other CDR datasets.

REFERENCES

[1] Cisco, "Global Mobile Data Traffic Forecast Update." 2013.

[2] United Nations, "Global Report on Human Settlements 2009." 2009.

[3] U.K. Paul, A.P. Subramanian, M.M. Buddhikot, S.R. Das, "Understanding Traffic Dynamic in Cellular Data Networks." *IEEE Infocom*, 2011.

[4] B. Wang, R.K.J. Liu, "Advances in cognitive radio networks: A survey." *IEEE Journal of Selected Topics in Signal Processing*, 5(1):5–23, 2011.

[5] K. Lee, J. Lee, Y. Yi, I. Rhee, S. Chong, "Mobile Data Offloading: How Much Can WiFi Deliver?" *IEEE/ACM Transactions on Networking*, 21(2):536–550, 2013.

[6] S. Kadloor, A. Raviraj, "Relay selection and power allocation in cooperative cellular networks." *IEEE Transactions on Wireless Communications*, 9(5):1676–1685, 2010.

[7] S. Liu, J. Wu, C.H. Koh, V.K.N. Lau, "A 25 Gb/s(/km2) urban wireless network beyond IMT-advanced." *IEEE Communication Magazine*, 49(2):122–129, 2011.

[8] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns", *Nature*, 453, 2008.

[9] A. Pawling, N. Chawla, G. Madey, "Anomaly detection in a mobile communication network", *Comput. Math. Organ. Theory*, 13(4):407-422, 2007.

[10] S. Isaacman, R. Becker, R. Cceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Ranges of human mobility in los angeles and new york", *IEEE MUCS*, 2011.

[11] S. Isaacman, R. Becker, R. Cceres, M. Martonosi, J. Rowland, A. Varshavsky and W. Willinger, "Human mobility modeling at metropolitan scales", *ACM MobiSys*, 2012.

[12] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility", *Science*, 327, 2010.

[13] R. A. Becker, R. Cceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, C. Volinsky, "A tale of one city: Using cellular network data for urban planning", *IEEE Pervasive Computing* 10(4):18–26, 2011.

[14] O. A. Abidogun, C. W. Omlin, "A self organizing maps model for outlier detection in call data from mobile telecommunication networks", *SATNAC*, 2004.

[15] S. Hoteit, S. Secci, Z. He, C. Ziemlicki, Z. Smoreda, C. Ratti, G. Pujolle, "Content consumption cartography of the paris urban region using cellular probe data", *UrbaNe*, 2012.

[16] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records", *Journal of Physics A: Mathematical and Theoretical*, 41(22), 2008.

[17] A. Vaccari, L. Liu, A. Biderman, C. Ratti, F. Pereira, J. Oliveirinha, A. Gerber, "A holistic framework for the study of urban traces and the profiling of urban processes and dynamics", *IEEE ITSC*, 2009.

[18] P. Paraskevopoulos, T. Dinh, Z. Dashdorj, T. Palpanas, L. Serafini, "Identification and Characterization of Human Behavior Patterns from Mobile Phone Data", *NetMob*, 2013.

[19] R. Pulselli, P. Ramono, C. Ratti, and E. Tiezzi, "Computing urban mobile landscapes through monitoring population density based on cellphone chatting", *Int. J. of Design and Nature and Ecodynamics*, 3, 2008.

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

[21] R. Xu,D. Wunsch, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[22] G. W. Milligan, M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, 50(2):159–179, 1985.

[23] Orange D4D challenge, http://www.d4d.orange.com.

[24] K. Pearson, "Notes on the History of Correlation", *Biometrika*, 13:25–45, 1920.