Mobile Demand Profiling for Cellular Cognitive Networking

Angelo Furno, Diala Naboulsi, Razvan Stanica, Marco Fiore

Mobile Demand Profiling for Cellular Cognitive Networking

Angelo Furno[1], Diala Naboulsi [1 2], Razvan Stanica[1], and Marco Fiore [3]

[1] *Univ. Lyon, INSA Lyon, Inria, CITI, F-69621 Villeurbanne, France*
[2] *CIISE, Concordia University, Montreal, QC, H3G2W1, Canada*
[3] *CNR-IEIIT, 10129 Torino, Italy*

In the next few years, mobile networks will undergo significant evolutions in order to accommodate the ever-growing load generated by increasingly pervasive smartphones and connected objects. Among those evolutions, cognitive networking upholds a more dynamic management of network resources that adapts to the significant spatiotemporal fluctuations of the mobile demand. Cognitive networking techniques root in the capability of mining large amounts of mobile traffic data collected in the network, so as to understand the current resource utilization in an automated manner. In this paper, we take a first step towards cellular cognitive networks by proposing a framework that analyzes mobile operator data, builds profiles of the typical demand, and identifies unusual situations in network-wide usages. We evaluate our framework on two real-world mobile traffic datasets, and show how it extracts from these a limited number of meaningful mobile demand profiles. In addition, the proposed framework singles out a large number of outlying behaviors in both case studies, which are mapped to social events or technical issues in the network.

Keywords: Cellular Networks, Cognitive Networking, Network Demand Profiling, Outliers Detection, Mobile Data Analysis

1 Introduction

Global mobile traffic is today growing at a dramatic pace. Over the last decade, its compound annual growth rate (CAGR) was higher than that recorded for Internet traffic during the surge of the World Wide Web at the turn of the millennium. Similar trends are expected to endure in the next years, with an anticipated 11-fold growth between 2014 and 2018 [2]. Facing the growth of the mobile demand is the foremost challenge for mobile operators, and the specifications for next-generation 5G cellular networks are tailored to meet it, putting forward a 1000-fold increase in wireless capacity, and a 100-fold increase in the number of supported connected devices [3]. To attain these goals, 5G architectures will not introduce incremental technological updates with respect to 3G and 4G networks ; rather, they will overhaul their legacy structures and operations.

Apart from major innovations in the radio technologies, 5G cellular systems will be designed for flexibility and re-configurability. New architectural paradigms, such as cloud radio access networking (C-RAN) and software-defined networking (SDN), will provide support for a dynamic, centralized management of resources both at the radio access and within the core infrastructure [4]. Solutions such as C-RAN and SDN will provide operational support to the realization of the *cognitive networking* paradigm in the context of cellular networks. Cognitive networking, first introduced in [5], extends the basic principle behind popular cognitive radios to the management of all network resources, paving the way for networked systems that self-manage through demand-aware functions for resource provisioning, optimization and troubleshooting. The cognitive networking approach is especially relevant in mobile networks such as cellular ones, where subscribers tend to consume resources in significantly different ways, depending on the time at which they access the network and on the location where they do so. As a result, the aggregated demand features very diverse macroscopic network utilization profiles over space and time, making an adaptive allocation of resources very beneficial in the economy of the system. Cellular cognitive networks will be able of perceiving such fluctuations in the user demand, and of reacting accordingly.

In order to realize the vision above, algorithmic solutions are needed that drive, in concert with the variations in the mobile demand, the establishment, modification, release and relocation of any type of resources in the network. This raises, in turn, the fundamental problem of understanding the mobile demand, and linking it to the resource management processes. In this paper, we tackle the very last problem above, and take a first step in the direction of designing *mobile traffic analytics* for the exploration of data collected within the cellular network and the inference of knowledge that is relevant to cognitive network management operations. Specifically, we propose a framework for the automated profiling of the mobile demand in large-scale cellular networks. Our framework runs on data gathered by the operator, is parameter-free, and allows constructing sensible categories of the demand that are associated to macroscopic spatiotemporal routines of the user population. As an interesting by-product, the framework can identify unusual behaviors in the demand, caused by events biasing the customary dynamics of a significant subscriber fraction.

We evaluate the effectiveness of the framework in two citywide case studies, leveraging large datasets collected by mobile operators and featuring over 300 million events each. We demonstrate that the proposed framework can successfully profile the mobile demands, and show how these latter can feature sensibly different spatiotemporal structures in our two case studies. Also, we detect through the framework a large number of outlying behaviors, which we show to have social or technical origins.

2 Related work

The analysis of mobile phone data has received significant attention over the last few years, as reviewed in [6].

Mobile traffic analyses and networking. Several works on human mobility leverage mobile phone data, usually presented under the form of Call Detail Records (CDRs), to characterize individual and population movements [7, 8], and predict them [9]. In these studies, observations are aggregated over time in order to draw conclusions on human mobility that can be useful to support network planning. 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 [10], for urban planning goals. In [11], subscribers' activity is clustered so as to distinguish categories of mobile users, and make a limited number of typical user profiles emerge. 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 [12]. However, such analysis considers only two days, known to represent typical and unusual network usages. Our objective is to start from large datasets, reporting on months of mobile traffic demand, and infer similarities among traffic profiles to categorize them.

In [13], the authors analyze the 2013 D4D Challenge [14] datasets to discover abnormal patterns in the communication flows of Ivory Coast population over a 5-month period. The proposed approach labels as outliers those time slots with a high number of base stations exhibiting an hourly traffic strongly deviating from its expectation. Differently from our solution, the approach heavily relies on arbitrarily defined parameters and thresholds.

Other recent works on country-scale CDR datasets aim at supporting network planning, operations and anomaly detection, by revealing the socio-economic structure of an area and capturing the influence of one region on another. Such a study has been recently carried out for the city of Milan in [15], on the datasets of the 2014 Telecom Italia Big Data Challenge [16]. By using spectral methods, the authors decompose time series of the aggregate cell phone activity into seasonal and residual components to distinguish between routine activities and deviations from them. The authors use this information to produce spatial clusters of base stations, which can be mapped to different land-use categories. Our proposed framework has a different objective than that in [15] : it aims at detecting the moments when parts (or all) of the network deviate from a normal behavior, rather than at characterizing the typical behavior of the demand at individual base stations.

Attaining our goal above requires a definition of similarity that allows placing network usages in the same

group. Previous works mainly focused on the total traffic volume when characterizing users' behaviors [17]. 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 [18], or by looking at each base station independently [19]. 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 [20] and [21] 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 only focus on the normalized volume, and do not consider any measure of similarity between traffic patterns nor provide a classification of network usage profiles.

Improving mobile networks with traffic analysis. Mobile traffic analytics has recently been used to evaluate and improve the performance of current mobile networks. In this sense, [22] studies the local mobile traffic peaks and proposes the tuning of the Radio Resource Control (RRC) protocol state machine in order to avoid performance degradation in presence of these special events. Understanding the spatiotemporal user behavior can also lead to traffic prediction models. In [23], the authors use such a model to reduce the energy consumption of user equipments by optimizing the functioning of the RRC state machine. Their proposed scheme saves more than 50% of the energy on the user side. Going beyond parameter tuning, [24] uses mobile traffic data to evaluate content caching at different levels of the cellular network architecture. The study shows that cache hit ratios between 27% and 33% can be achieved when content is stored within the core network.

All these studies focus on the current cellular architecture, and mobile phone data are rarely used to evaluate novel architectures and mechanisms for next-generation 5G cellular networks. Practically, when using such massive datasets to study evolved networking architectures, the focus is generally on the integration of device-to-device (D2D) communication in mobile networks [22, 25]. At the same time, topics such as C-RAN [26], green networking [27], and machine-to-machine (M2M) support [28] gain popularity in the context of 5G architectures. While still on the drawing board, these technologies mark the movement from a network with a fixed number of resources towards a flexible, adaptive architecture, where resources are added only when necessary and removed when no longer needed. The approach in this case is to design mechanisms capable of modifying the type and quantity of available resources at times when the user demand changes. Our work is complementary to these solutions, profiling the demand and allowing the operators to decide when and where to trigger the different adaptive mechanisms.

3 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 any type of available mobile traffic data. As the name suggests, a snapshot is a representation of the load generated by mobile users on the access network during fixed-size time intervals. We do not impose any constraint on the way snapshots are defined, e.g., 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 (voice calls, short text messages, Internet-based applications, etc.). In the following, we will denote as $\mathbb{T}$ the set of snapshots that we aim at analyzing with our framework. Each snapshot will be uniquely identified by the first instant of the time interval it refers to. Similarly, $\mathbb{Z}$ will indicate the set of geographical areas over which traffic volumes are aggregated. The choice of $\mathbb{T}$ and $\mathbb{Z}$ may depend on the level of detail of the available mobile traffic data or on the target of the study, yet our framework is general enough to accommodate any definition of such sets. We will provide practical examples of snapshot definitions when introducing the datasets employed for our performance evaluation, in Sec. 4.

Once snapshots are defined and extracted from the mobile traffic data, 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 classifying additional usage profiles into the categories above. The different phases are detailed in the rest of this section.

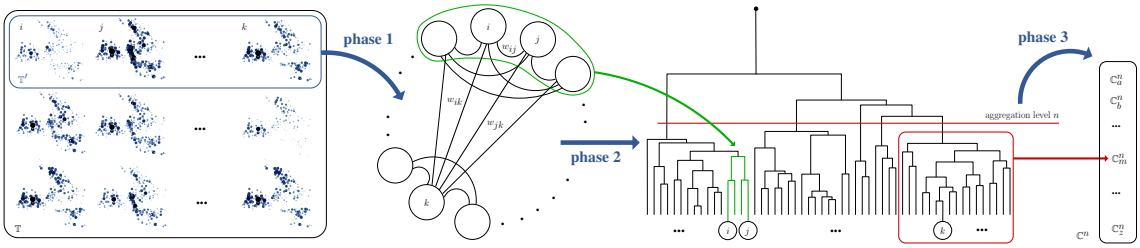


FIGURE 1: Workflow of the framework for the definition of categories of network usage profiles. Phase 1 : construction of the snapshot graph from snapshots (portrayed here as geographical plots of the mobile traffic volume) in the training set $\mathbb{T}'$. Phase 2 : iterative aggregation of graph vertices 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.

3.1 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 mobile traffic data. For instance, 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.

One option is to directly use the selected snapshots in the following steps of the framework. Another option is to pre-process them, in order to remove potential biases that can be introduced by outlying behaviors present in the training set. To that end, one approach is to perform a filtering step to aggregate those snapshots in $\mathbb{T}'$ that are separated by a fixed time interval. As an example, by using a 1-week distance, all the snapshots in $\mathbb{T}'$ referring to Thursdays at 9 :00 are merged into a *median Thursday at 9 :00* snapshot [†]. Iterating over $\mathbb{T}'$, it is possible to generate a synthetic training set $\overline{\mathbb{T}}'$, characterized by the same temporal and spatial granularities as $\mathbb{T}'$ and representative of the *median week*. Such synthetic set can be used in place of $\mathbb{T}'$ in the following. The latter is the approach we adopt for the case studies in Sec. 4.

Snapshots in $\mathbb{T}'$ (or in the equivalent median week) are then mapped to the vertices of an undirected weighted graph $G(\mathbb{T}', \mathbb{E})$ that we dub *snapshot graph* (see phase 1 in Fig. 1). In the definition above, $\mathbb{E} = \{e_{ij} \,|\, i, j \in \mathbb{T}', i \neq j\}$ is the set of edges e_{ij} between any two snapshots i and j of the training set $\mathbb{T}'$: thus, the snapshot graph is a clique. 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 [‡] observed in the geographical area $z \in \mathbb{Z}$.

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 (e.g., [22]).

However, while it permits to identify large positive or negative variations in mobile traffic, this metric does not account for spatial diversity. Thus, we introduce a *traffic volume similarity* measure $\mathcal{V}$ that accounts for geographical sub-areas when computing traffic volume variations between two snapshots i and j. Formally :

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

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.

†. Aggregation is based on the median, as it is less sensitive to outlying behaviors than other metrics, e.g., the average.

‡. As previously stated, our definition of mobile traffic volume is general. Depending on the available mobile traffic data 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 services.

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 captures how mobile traffic is divided among different areas. The weight between two snapshots i and j is then :

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

Here, 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 in the case of $\mathcal{V}$. This allows capturing how the traffic is distributed over the region, independently of its absolute volume.

In our case studies in Sec. 4, we use both $\mathcal{V}$ and $\mathcal{D}$ as snapshot similarity measures (i.e. w_{ij}), as they are complementary in the identification of network usage profiles. This implies that one snapshot graph is built for each measure, and that the next phases are performed separately on the two graphs.

3.2 Snapshot aggregation

The snapshot graph is used in the second phase as a base for the definition of a set of potential categories of network usage profiles. Here, the goal is to identify all meaningful 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 (see phase 2 in Fig. 1).

We adopt the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [29] – 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. At every iteration, the algorithm merges the two clusters that share the strongest tie : this means aggregating the groups of snapshots that yield the highest similarity in terms of network usage profiles. Iterations continue until all nodes are grouped into one cluster.

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

$$d_{kh}^n = \frac{1}{|\mathbb{C}_k^n| \cdot |\mathbb{C}_h^n|} \sum_{i\in\mathbb{C}_k^n, j\in\mathbb{C}_h^n} \frac{1}{w_{ij}}.$$

Once such a value has been computed for all pairs, the two clusters $\mathbb{C}_k^n$ and $\mathbb{C}_h^n$ having the smallest average distance are joined into a new cluster $\mathbb{C}_m^{n+1}$, 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.

3.3 Network usage profile categories

The dendrogram generated by UPGMA represents a full family of clusterings, as each level in the dendrogram maps to one possible partitioning of snapshots. One must then choose the best clustering, i.e., dendrogram level : the resulting clusters will become our network usage profile categories (see phase 3 in Fig. 1).

Many criteria, or stopping rules, have been proposed to automatically detect the best clustering in dendrogram structures : however, there is no clear winner among previously proposed stopping rules, which may in fact return inconsistent results [30]. In order to achieve a dependable result, we introduce an original criterion, named *top-k index*, which aggregates the output of a set $\mathbb{S}$ of base stopping rules. We let each stopping rule $s \in \mathbb{S}$ select its k best clusterings over the dendrogram, and then rank clusterings based on how many rules picked them in their top-k list [§]. The best clustering is that attaining the highest rank in the top-k index. Formally, for the dendogram level $n \in [2, |\mathbb{T}'|]$:

§. The k parameter controls the number of suggestions allowed for each base stopping rule. A smaller k increases the precision of the aggregate index, but generates a higher number of approximately equivalently good recommendations. From our experiments, we find that $k = 10$ works well with all our reference datasets.

$$top\text{-}k^n = \sum_{s \in \mathbb{S}} \left(1 - \frac{rank_s^n}{k}\right),$$

where $rank_s^n \in [0,k)$ is the position of the clustering corresponding to level n in the ranking returned by the base stopping rule $s \in \mathbb{S}$.

We referred to the extensive survey in [30] to make our choice of base stopping rules in $\mathbb{S}$. Specifically, we implemented and employed seven top-ranked, popular rules among those proposed in the literature. They are the Calinski-Harabasz, Beale, Duda-Hart, C, Hartigan, Krzanowski-Lai and Silhouette indices, briefly described in the following. Still, the top-k index is extensible to any number and different selection of stopping rules.

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 [31] :

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

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

There, $\bar{c}_k^n$ is the *center* of cluster $\mathbb{C}_k^n$, i.e., 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 $\bar{s}$ is a synthetic snapshot representing the center of all snapshots in the training set $\mathbb{T}' = \bigcup_{\mathbb{C}_k^n \in \mathbb{C}^n} \mathbb{C}_k^n$.

The 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 dispersion 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 $\bar{c}_k^n$. Clearly, B^n and P^n respectively decrease and increase as n grows : the second factor, iversely proportional to the number of clusters $|\mathbb{C}^n|$, compensates this, allowing for a fair comparison across different aggregation levels.

Overall, the CH index compares the distance among clusters B^n to the level of internal scattering 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 [32] :

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

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

where h is the number of observed variables on the clustered objects (i.e., $h = |\mathbb{Z}|$ in our case). 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 $F\left(h, (|\mathbb{C}_m^{n+1}| - 2)h\right)$ at a significance level of 5%. The null hypothesis that the clustering quality at level $n+1$ is better than that at level n is rejected if $F^n > F_{crit}^n$. Therefore, the stopping rule suggests to select those dendrogram levels with higher values for the difference $F^n - F_{crit}^n$.

C index. The C-index is based on the computation of all the intra-cluster distances for a given partitioning of a dataset. It is computed as [33] :

$$C^n = \frac{S^n - S_{min}^n}{S_{max}^n - S_{min}^n}, \text{ with } S^n = \sum_{\mathbb{C}_k^n \in \mathbb{C}^n} \sum_{i,j \in \mathbb{C}_k^n, i \neq j} \frac{1}{w_{i,j}}.$$

S^n is the sum of the distances between all pairs of objects from the same cluster at level-n aggregation. By denoting as r the total number of those pairs, S^n_{min} (respectively, S^n_{max}) is the sum of the r smallest (respectively, largest) distances, if all possible pairs of objects are considered. The value of n which minimizes the C-index is considered as a good choice for the best clustering.

Duda-Hart index. This stopping rule is based on the evaluation of the ratio between the sum of squared errors within two clusters and the squared errors when using only one cluster. The index allows deciding if it is preferable to split one cluster (i.e., $\mathbb{C}^n_m$) in two (i.e., $\mathbb{C}^{n-1}_k$ and $\mathbb{C}^{n-1}_l$), thus reducing the level of aggregation, or to stop at the current cluster configuration (i.e., level n). It is local, in the sense that the only information it uses comes from the single cluster being splitted at the current n-level. The stopping rule requires the evaluation of a critical value (i.e., DH_c) and the comparison of the DH index with it. Formally [34] :

$$DH = \frac{P^{n-1}_k + P^{n-1}_l}{P^n_m}, \text{ and } DH_c = 1 - \frac{2}{\pi h} - K\sqrt{\frac{2\left(1-\frac{8}{\pi^2 h}\right)}{|\mathbb{C}^n_m| h}},$$

where h represents the number of observed variables and $K = 3.2$ is a typical value used in DH_c formula. The stopping rule suggests to consider the lowest number of clusters for which $DH \geq DH_c$.

Silhouette index. This index is defined as [35] :

$$silhouette^n = \sum_{i=1}^{|\mathbb{T}'|} \frac{S^n_i}{|\mathbb{T}'|}, \text{ with } S^n_i = \frac{b^n_i - a^n_i}{\max\{a^n_i, b^n_i\}},$$

where a^n_i represents the average distance of snapshot i from all the other snapshots within the same cluster and b^n_i is the lowest average distance of i from any other cluster, of which i is not a member. Being $\mathbb{C}^n_r$ the cluster to which snapshot i belongs, they are defined as follows :

$$a^n_i = \frac{\sum_{j \in \mathbb{C}^n_r, j \neq i} (1/w_{i,j})}{|\mathbb{C}^n_r| - 1} \quad \text{and} \quad b^n_i = \min_{k \neq r} \frac{\sum_{j \in \mathbb{C}^n_k} (1/w_{i,j})}{|\mathbb{C}^n_k|}.$$

The level n maximizing $silhouette^n$ is then selected.

Hartigan index. By re-using the definition of P^n from the CH index to represent the intra-cluster distance, the Hartigan index is defined as follows [36] :

$$hartigan^n = \left(\frac{P^n}{P^{n+1}} - 1\right)\left(|\mathbb{T}'| - n - 1\right)$$

The level n maximizing $hartigan^n$ is then selected.

Krzanowski and Lai index. The KL index uses the intra-cluster distance P^n to determine the best number of clusters. It is defined as follows [37] :

$$KL^n = |\Delta_n / \Delta_{n+1}|, \text{ with } \Delta_n = (n-1)^{2/p} P^{n-1} - (n)^{2/p} P^n.$$

The level n maximizing KL^n is then selected.

3.4 Snapshot classification

Stopping rules allow us to define the aggregation level at which clusters of snapshots show the best trade-off 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 [29]. While k-means is commonly known as a partitional clustering algorithm, here we use a simplified version of the k-means algorithm as a classification technique. As a clustering algorithm, k-means assigns a set of objects to a predetermined number k of clusters, by starting from an arbitrary initial assignment of all objects to the k clusters. To that end, it performs multiple iterations of an assignment step, based on the computation of the lowest distance between the object and the cluster centroids, until a convergence criterion is met.

In our case, instead, we consider the categories defined at the end of the UPGMA training phase above as the initial partitioning of data for k-means, and perform one single iteration of the assignment step to classify each snapshot in $\mathbb{T} \setminus \mathbb{T}'$. The distance measure used in our simplified k-means implementation is the average distance $1/w_{i,\bar{c}_k^n}$ between the yet-unclassified snapshot i and each cluster $\mathbb{C}_k^n \in \mathbb{C}^n$. The algorithm then assigns the snapshot to the category for which such a measure is the smallest. We call the resulting category the *actual class* of snapshot i, i.e., $\mathbb{C}_{act}(i)$.

If the training process is performed over the median week $\overline{\mathbb{T}}'$ (see Sec. 3.1), we can denote the *expected class* $\mathbb{C}_{exp}(i)$ for snapshot i as follows. Let us describe the generic snapshot i by the corresponding day of the week W (Monday to Sunday), time t. Then, we consider the median snapshot $x \in \overline{\mathbb{T}}'$, representing the typical behavior for day of the week W and time t, and select its class from the UPGMA output, i.e., $\mathbb{C}_{W,t}^n \in \mathbb{C}^n$. The latter is the expected class for snapshot i, i.e., $\mathbb{C}_{exp}(i) = \mathbb{C}_{W,t}^n$. As an example, the expected class for Thursday, 20$^{\text{th}}$ November at 20 :00 will be the class of the median Thursday at 20 :00 from $\overline{\mathbb{T}}'$.

Two types of outlying behaviors, deviating from usual routines, can be detected :

1. if $\mathbb{C}_{act}(i) \neq \mathbb{C}_{exp}(i)$, snapshot i is closer to a different profile than the one of the corresponding median snapshot. In other words, i diverges from the typical behavior in such a way that the associated mobile traffic data resembles those of another profile from the cluster set, according to the selected similarity measure.
2. if $\mathbb{C}_{act}(i) = \mathbb{C}_{exp}(i)$, distance $1/w_{i,\bar{c}_{exp}}$ is considered as an indication of outlying behavior. Even though the closest cluster matches the expected profile, i can be significantly far from the cluster centroid and, therefore, not well described by the associated class.

In Sec. 5, we will use two different graphical notations to pinpoint both kinds of outliers.

4 Datasets

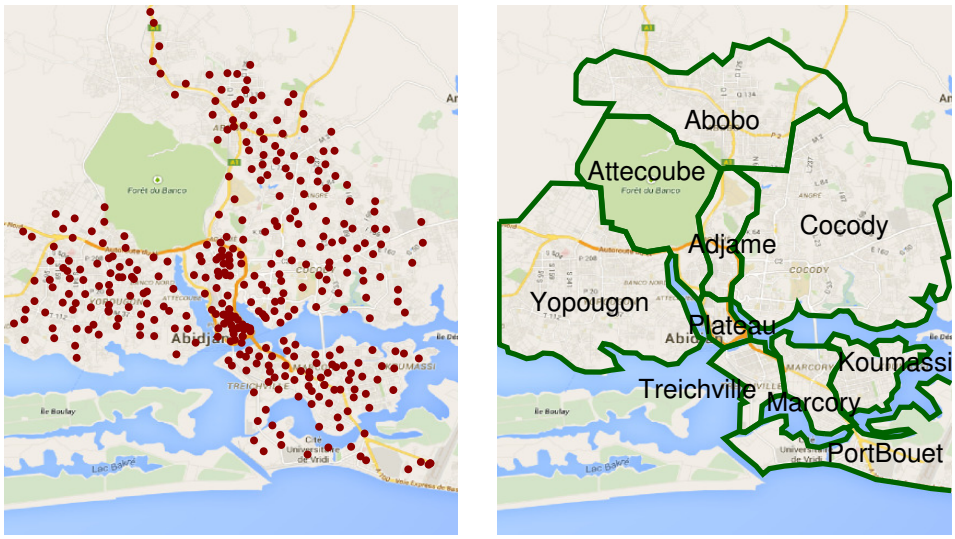
We evaluated our framework on two case studies. The first one maps to a dataset provided by Orange within the context of the 2013 D4D Challenge [14]. The second one refers to data published as part of the Big Data Challenge organized by Telecom Italia in 2014 [16].

4.1 D4D Orange Dataset

Our first use-case dataset is based on anonymized CDRs of 5 million Orange customers in Ivory Coast, and describes the mobile traffic volume in terms of the number of voice calls exchanged between any two base stations of the operator in the whole country, aggregated on a hourly basis. The information covers over 5 months, from December 5$^{\text{th}}$, 2011 to April 22$^{\text{nd}}$, 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 in Abidjan, i.e., 364 base stations with the deployment in Fig. 2a.

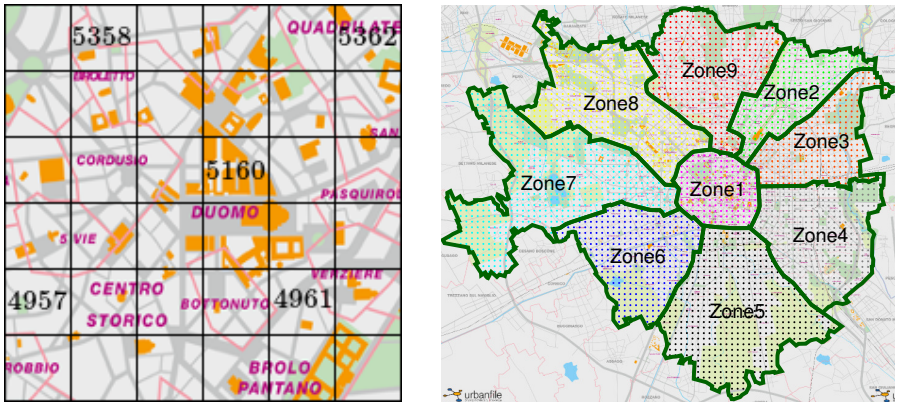
Coherently with the D4D dataset granularity, we consider snapshots to aggregate information at each hour. Thus, our set $\mathbb{T}$ contains over 3600 snapshots, each describing the network usage over a specific hour. The set of geographical areas $\mathbb{Z}$ over which traffic volumes are aggregated is mapped to the set of *communes* of the city, shown in Fig. 2b. While we acknowledge that different options might be envisioned, this spatial aggregation is intuitive, as it is based on topological criteria. Additionally, it has been proved in [38] that the administrative division of Abidjan corresponds to similar typical behaviors of the different base stations described in the D4D dataset. The available dataset only contains voice call volumes, and therefore all our results refer to voice traffic.

During our analysis, we noted that the information regarding the 364 Abidjan antennas is not always present for the entire observation period. As reported in the supplemental material to this paper, we found out that three main behaviors can be identified and explained by different collection phases ; during each period, voice call traffic was recorded for a different (but not disjoint) subset of base stations. Additionally, we identified a fourth scenario including snapshots affected by technical problems encountered by the operator in data collection and occasional power grid failures in Abidjan. We eliminate from $\mathbb{T}$ and do not consider in our analysis the snapshots in the last scenario above.



(a) (b)

FIGURE 2: (a) Base station deployment and (b) communes in Abidjan.



(a) (b)

FIGURE 3: (a) Division of Milan as a grid zoomed at Milan's Duomo [16]. (b) Administrative zoning of Milan, cells centers.

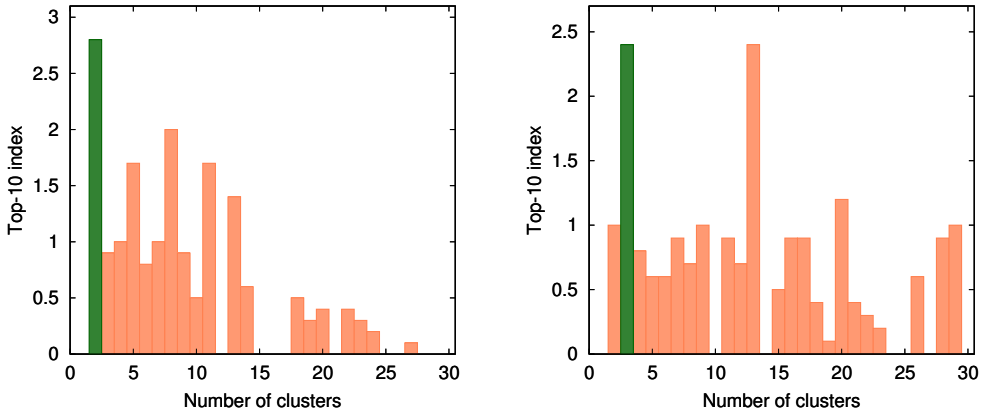
It is noteworthy that our decision of analyzing snapshots collected in different periods implies that two different snapshots i and $j \in \mathbb{T}$ can contain a different number of base stations. In order to fairly compute the similarity w_{ij} of snapshots i and j with a different number of base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

4.2 Telecom Italia Big Data Challenge Dataset

The dataset provided by Telecom Italia is based on a tessellation of the surface of the city of Milan in cells, see Fig. 3a. These cells represent the highest spatial granularity at which mobile traffic measurements are provided by the Italian operator, and, differently from the Abidjan scenario, they give no information about the deployment of actual base stations in the area. Each square has a $235m \times 235m$ size, and the grid is composed of 10,000 squares.

To define the geographical areas $\mathbb{Z}$ for traffic volumes aggregation by our framework, we have used the current administrative subdivision of Milan, made of nine *decentralization zones*. Fig. 3b shows these zones. The administrative zones division includes 3,339 cells of the original Milan grid.

The Telecom Italia dataset analyzed in this paper reports on subscribers' communication activity in terms of received and sent text messages, incoming and outgoing calls and Internet data usage. This data has been



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

FIGURE 4: Abidjan. Top-10 score versus the number of clusters (i.e., the dendrogram aggregation level) for the median week training set.

extracted from the operator's CDRs and aggregated with respect to each cell of the Milan grid, covering the time period from November 1st, 2013 to January 1st, 2014. Measurements are temporally aggregated in time slots of ten minutes. We further aggregated the available data to one-hour bins, similar to the Abidjan use case. Our final dataset contains 1,488 snapshots.

As detailed in the supplemental material, the Milan dataset is more reliable than the Abidjan one, and thus it does not require the cleansing process in Sec. 4.1.

5 Results

We ran the framework of Sec. 3 on the datasets of Sec. 4. The outcome for Abidjan and Milan are presented separately.

5.1 Case Study 1 : Abidjan

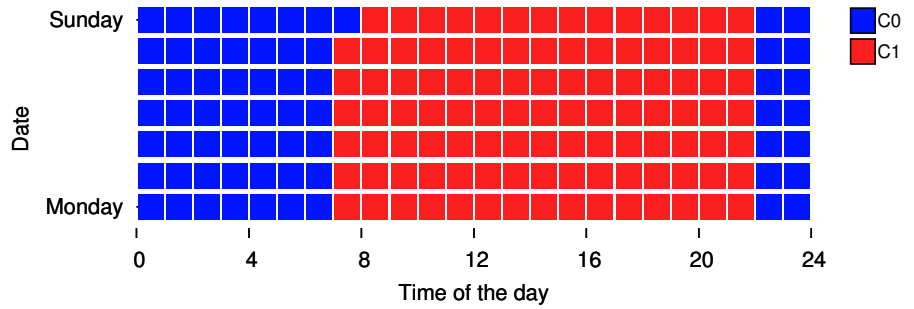
We first present the results in the Abidjan scenario. As mentioned in Sec. 4.1, the mobile traffic data concerns exclusively voice, i.e., calling activity in this case.

5.1.1 Call profile categories

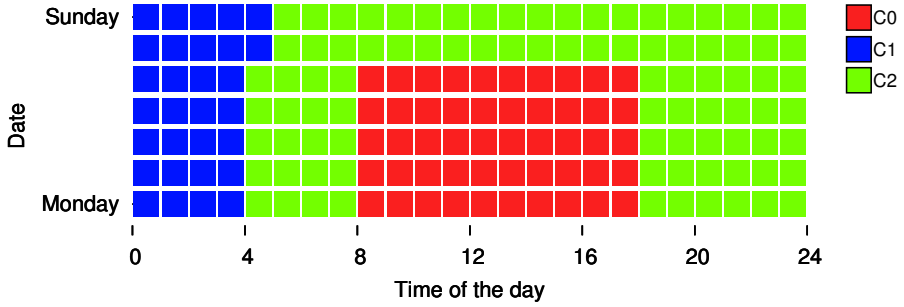
As explained in Sec. 3, our framework requires a training set $\mathbb{T}' \subseteq \mathbb{T}$ from which to derive the different categories of network usage profiles. We trained the framework on the *median week* training set $\overline{\mathbb{T}}'$, computed as described in Sec. 3.1, and adopting a leave-one-out approach on a weekly basis for the classification of snapshots in $\mathbb{T} \setminus \mathbb{T}'$. The plots in Fig. 4 portray the evolution of the aggregate top-k clustering index, introduced in Sec. 3.3, versus the number of clusters n. Fig. 4a refers to the dendrogram obtained under the traffic volume similarity measure $\mathcal{V}$, and Fig. 4b to that obtained when using $\mathcal{D}$. The index recommends *two* clusters for $\mathcal{V}$ and *three* clusters for $\mathcal{D}$ as the best aggregation of snapshots in $\mathbb{T}'$, since it reaches its maximum at these values ¶.

We present the structure of the categories found over the training dataset for $\mathcal{V}$ and $\mathcal{D}$ in Fig. 5a and Fig. 5b, 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 7 :00 and 22 :00. More interestingly, for the case of the $\mathcal{D}$ metric, we can observe that the snapshots of the training set are divided into three clusters. The first category includes the snapshots of night hours, between 0 :00 and 4 :00, characterized by the lowest activity in the city center, and a high traffic concentration in residential (i.e., firstly, the wealthy area of Cocody and, secondly, the working-class populated Yopougon) and industrial (e.g., Cocody, Marcory) areas. The second category includes daytime snapshots from the weekdays, i.e., hours between 8 :00 and 18 :00, Monday to Friday : these snapshots show higher mobile

¶. To tie-break identical score situations, we select the lowest number of clusters, since wider categories are typically more reliable.



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



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

FIGURE 5: Mobile traffic profile categories defined on the training set $\overline{\mathbb{T}}'$. Each square represents one snapshot, and colors map to categories.

traffic in the office and university areas. The third category contains snapshots from weekend days, as well as from early morning (4 :00–8 :00) and evening (18 :00–24 :00) hours of weekdays : the corresponding network usage is that of a high traffic concentration in the residential areas (i.e., firstly, Yopougon and, secondly, Cocody and Abobo), and some low activity in the city center.

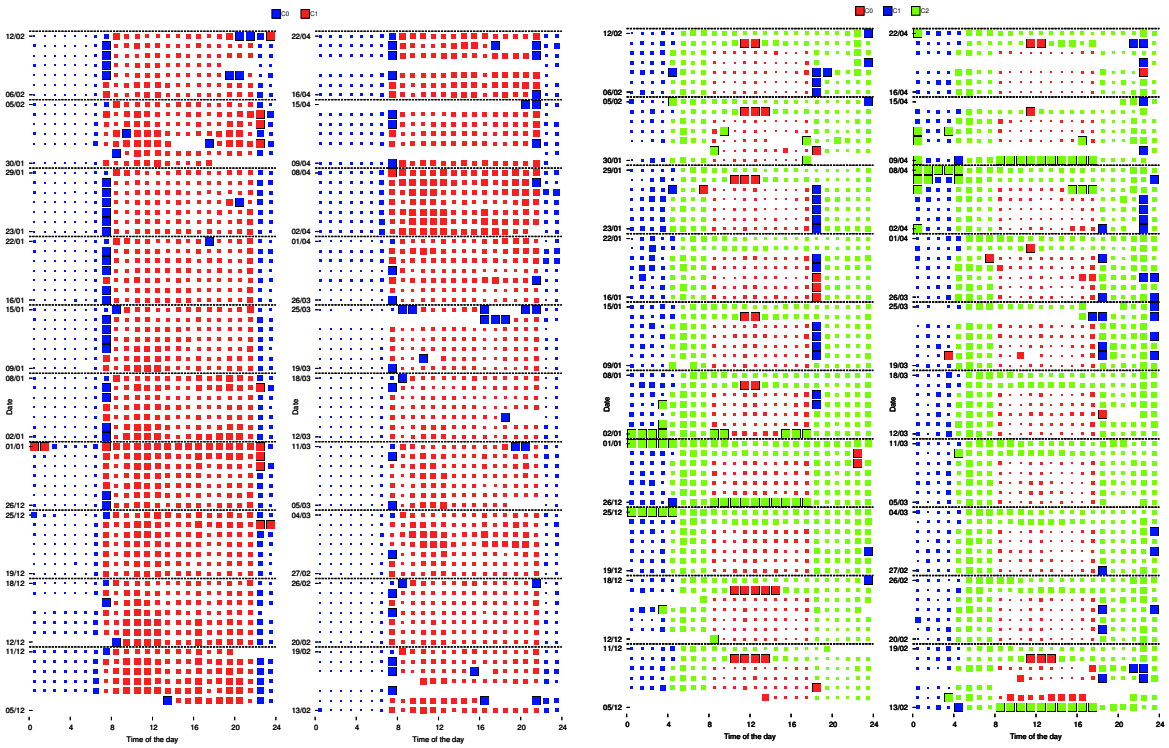
5.1.2 Call profile classification

Once categories have been defined on snapshots of the training set, all of the snapshots in $\mathbb{T} \setminus \mathbb{T}'$ can be classified accordingly, as described in Sec. 3.4. We plot in Fig. 6 the classification of the whole 5-month dataset, with respect to the categories defined under the measures $\mathcal{V}$ and $\mathcal{D}$.

The plots have the same format as those reported in Fig. 5. For each snapshot, they show with a different color the class it belongs to. Empty squares are snapshots that were filtered out from the dataset, as explained in Sec. 4.1. Additionally, the plots include two graphical representations for detecting outlying behaviors, as described in Sec. 3.4. First, snapshots that fall in a different category from the expected one are represented with big squares, surrounded by a black border (e.g., January 1st at 0 :00 – 2 :00 in Fig. 6a). Second, snapshots classified in the expected cluster are represented with squares whose size is proportional to the normalized average distance from the expected class. E.g., January 1st at 7 :00 – 22 :00 has a highly outlying behavior with respect to the expected cluster C1 in Fig. 6a [‖]. As a rule of thumb, the smaller is the square associated to a snapshot, the closer is the network usage to what expected.

We can observe in both Fig. 6a and Fig. 6b that the categories retain their initial structure, as most snapshots in the 5-month dataset are classified in what can be considered as their typical category. However, some snapshots join categories that differ from those they supposedly belong to, i.e., yield unusual network usage profiles. These are outliers, which we discuss in detail next.

‖. The average distances from the expected cluster have been normalized and saturated both with respect to sure outliers (e.g., January 1st at 00 :00) and with respect to the 95th percentile of all the distances, showing very similar results. Figures refer to the latter approach.



(a) Similarity measure $\mathcal{V}$.

(b) Similarity measure $\mathcal{D}$.

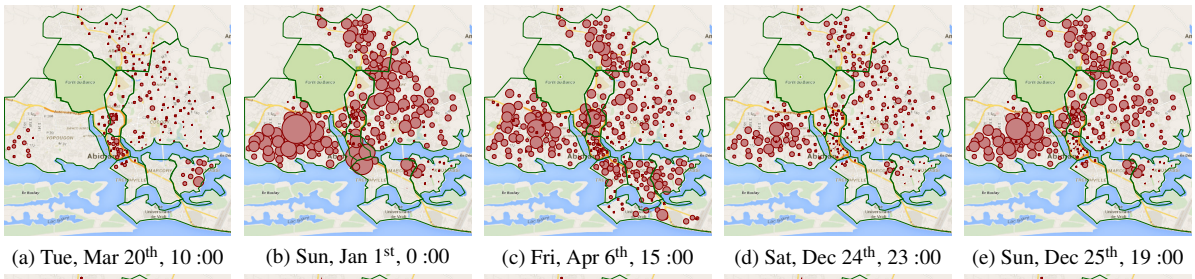
FIGURE 6: Classification of the 5-month mobile traffic data in Abidjan. (a) Categories obtained with $\mathcal{V}$. (b) Categories obtained with $\mathcal{D}$.

5.1.3 Call profile outliers

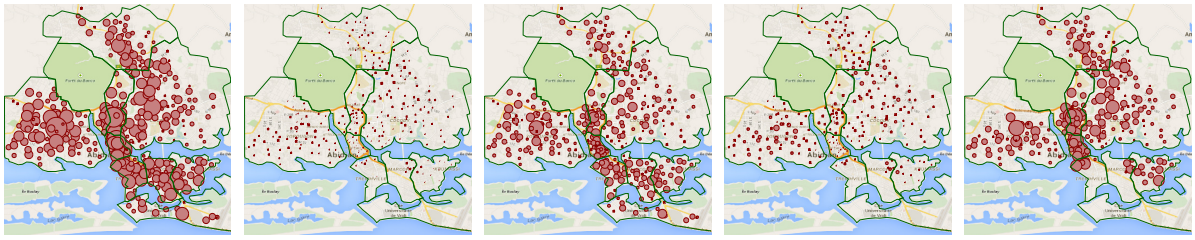
Focusing on the two categories obtained with the similarity measure $\mathcal{V}$, we observe that some snapshots at day time hours, such as 10 :00 and 16 :00, join unexpectedly the low-activity category in Fig. 5a. Clearly, these snapshots present outlying behaviors, in terms of number of calls. Some 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 as described in Sec. 4.1. 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 20$^{\text{th}}$ at 10 :00, whose call volume is portrayed in Fig. 7a. 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 3$^{\text{rd}}$ at 10 :00, in Fig. 7f, we can note that an important number of antennas is missing from the dataset on March 20$^{\text{th}}$. We consider this result as a demonstration that an automated framework can identify unusual network behaviors at a much finer granularity than a statistical analysis on aggregate data.

Concerning the outliers falling in cluster C1 in Fig. 6a, i.e., low-activity hours showing an uncommon surge in mobile traffic, these are mostly found around 22 :00, and are related to special events. This is the case, e.g., of New Year's Eve, portrayed in Fig. 7b, when people are calling much more than on a typical Sunday at the same time, e.g., January 8$^{\text{th}}$, in Fig. 7g. On the same principle, we could detect outliers on Christmas' Eve and right after the end of the final football game of the Africa Cup of Nations.

As far as the classification based on the $\mathcal{D}$ measure is concerned, in Fig. 6b we can observe multiple situations where snapshots diverge from the expected behavior. See, e.g., Friday, April 6$^{\text{th}}$ at 15 :00, depicted in Fig. 7c, with respect to a typical behavior portrayed in Fig. 7h, representing Friday, April 20$^{\text{th}}$ 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 C2 : it shows an increase in call



(a) Tue, Mar 20th, 10 :00 (b) Sun, Jan 1st, 0 :00 (c) Fri, Apr 6th, 15 :00 (d) Sat, Dec 24th, 23 :00 (e) Sun, Dec 25th, 19 :00



(f) Tue, Apr 3rd, 10 :00 (g) Sun, Jan 8th, 0 :00 (h) Fri, Apr 20th, 15 :00 (i) Sat, Dec 17th, 23 :00 (j) Mon, Feb 6th, 09 :00

FIGURE 7: Call volumes in Abidjan for different snapshots. In each plot, one base station maps to a dot, whose size is proportional to the traffic volume.

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 C2 : 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 C0, like several Saturdays between 11 :00 – 13 :00, can be related to events (e.g., Hilary Clinton and Kofi Annan's visit on January the 7th, "The Birth of the Prophet" holiday celebration on February the 4th) 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 C1 reflect a reduced level of calls in the network.

Another important result concerning $\mathcal{D}$ is the fact that snapshots with a different call volume, but with a similar traffic distribution, are assigned to the same category. For example, the snapshots presented in Fig. 7d and Fig. 7i both belong to C3, showing that, although people are making more calls on Saturday, December 24th at 23 :00, that 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 are placed in different clusters based on the $\mathcal{V}$ metric : in that case, December 24th at 23 :00 is considered an outlying behavior due to the increased traffic volume. We also detected cases where snapshots belonging to the same category based on $\mathcal{V}$ are 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. 7e and Fig. 7j.

Finally, some snapshots are classified in the expected category but are significantly far from the typical behavior of that class. We recall that, in Fig. 6, this is represented by means of bigger squares without a black border. For example, when considering $\mathcal{V}$ similarity, such kind of outlying behavior interests all the daytime snapshots related to January, 1st, most of those related to January, 2nd, and many of the daytime snapshots during the Easter week. Regarding $\mathcal{D}$ similarity, we can note that evening hours on January 1st and 2nd are quite far from the typical behavior of class C2. Similarly, the snapshots related to 14th February from 9 :00 to 17 :00 are distant from C0. Most of these outliers can be easily explained by considering their proximity to public holidays, when business and school activities are reduced.

Regarding February, 14th, instead, we may notice that on the same day many unreliable snapshots (from 17 :00 to 21 :00) are present. Some network problems, i.e., lots of base stations missing in Yopougon from 9 :00 to 17 :00, anticipate the final network outage from 17 :00 to 20 :00, when most of the base stations are missing in almost all communes.

In Tab. 1, we present a partial list of outliers detected by the classification obtained with measures $\mathcal{V}$ and $\mathcal{D}$, that we were able to relate to special events. The table also reports the category where each outlier was

TABLE 1: List of outlying snapshots, according to the classification for Abidjan with $\mathcal{V}$ and $\mathcal{D}$, respectively.

Date	*Similarity measure*	*Actual category*	*Expected category*	*Event*
Tue, Dec 6th, 13 :00 ; Tue, Feb 14th, 16 :00, 21 :00 ; ...	$\mathcal{V}$	C0	C1	Outliers before and/or after sequences of non-reliable unclassified snapshots, due to network outages
Sat, Dec 24th, 22 :00–0 :00	$\mathcal{V}$	C1	C0	Christmas' Eve
Sun, Dec 25th, 0 :00 ; 8 :00	$\mathcal{V}$	C0 (high distance)	C0	Christmas Day
Sun, Jan 1st, 0 :00–2 :00	$\mathcal{V}$	C1	C0	New Year's Eve
Sun, Jan 1st, 9 :00–22 :00	$\mathcal{V}$	C1 (high distance)	C1	First day of the year.
Wed, Feb 8th, 19 :00–21 :00	$\mathcal{V}$	C0	C1	Africa Cup of Nations semi-final. Much less traffic than usual
Sun, Feb 12th, 20 :00–22 :00	$\mathcal{V}$	C0	C1	Africa Cup of Nations final. Much less traffic than usual during the match
Sun, Feb 12th, 23 :00–0 :00	$\mathcal{V}$	C1	C0	Africa Cup of Nations final. Higher traffic than usual after the match
Sun, Dec 25th, 0 :00–5 :00	$\mathcal{D}$	C1	C2	Christmas night
Mon, Dec 26th, 8 :00–18 :00	$\mathcal{D}$	C2	C0	Public holiday compensating the Sunday Christmas
Fri, Dec 30th ; Sat 31th, 22 :00–23 :00	$\mathcal{D}$	C0	C2	"Perles des lumières" new year celebration in city center
Mon, Apr 9th, 8 :00–18 :00	$\mathcal{D}$	C2	C0	Easter Monday
Fri, Apr 6th, 15 :00–18 :00	$\mathcal{D}$	C2	C0	Good Friday
Wed, Dec 7th, 18 :00	$\mathcal{D}$	C0	C2	Celebration of the first Ivory Coast president in city center
Wed, Feb 8th, 18 :00–20 :00	$\mathcal{D}$	C1	C2	Africa Cup of Nations, semi-final match
Sun, Feb 12th, 23 :00	$\mathcal{D}$	C1	C2	Africa Cup of Nations final match
Mon, Feb 13th, 8 :00–18 :00	$\mathcal{D}$	C2	C0	Public holiday, celebration of Ivory Coast results in Africa Cup of Nations

detected, the one where it was expected to end, and the social reason behind the outlying behavior. Finally, we underscore that many more outliers were identified by our framework, as shown in Fig. 6 : however, a comprehensive discussion is impractical, due to space constraints and limited side information about minor events that could have caused many outliers. We refer the interested Reader to the sigificant additional outliers listed in Tab. 1 of the supplemental material.

5.2 Case Study 2 : Milan

In the Milan use case, we exploited our framework to analyze the records related to the different kinds of subscriber activity, i.e., incoming and outcoming calls, incoming and outgoing text messages, and Internet data. Due to space limitations, here we present only the results for incoming calls (call-in in the following), which represent an interesting sample of the mobile traffic activity in the urban area of Milan. We also provide in Tab. 2 a description of some outliers, retrieved by our framework, in relation to the other kinds of activity. The interested reader can access the supplemental material, provided with this manuscript, for a comprehensive presentation of the output of our framework when fed with other types of mobile traffic activity.

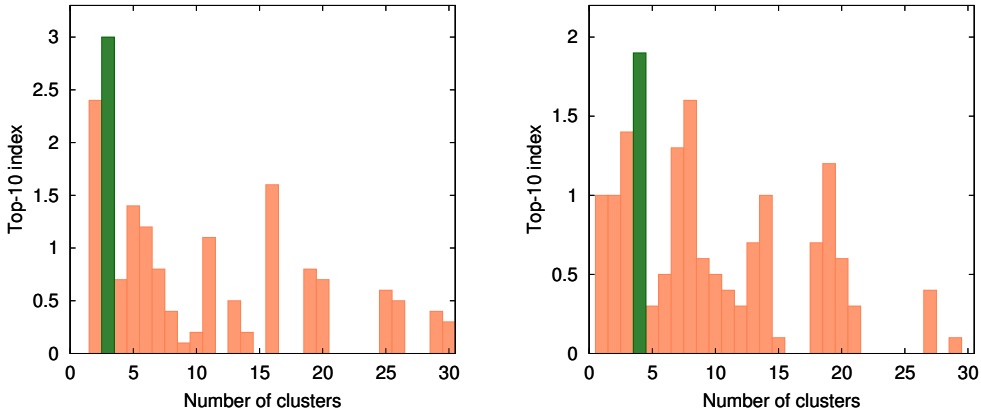
5.2.1 Call-in profile categories

As we did for Abidjan, we trained our framework on the median week, computed from the whole 2-month call-in dataset. Fig. 8a and Fig. 8b report on the top-10 scores versus the number of clusters, computed with the $\mathcal{V}$ and $\mathcal{D}$ similarity measures. They recommend to select *three* and *four* categories, respectively.

The three categories identified by the $\mathcal{V}$ measure, in Fig. 9a, are associated with very neat behaviors in relation to call-in volumes. Class C2 groups high-traffic snapshots, mainly related to working and homecoming time, i.e., Monday to Friday from 9 :00 to 20 :00. Class C1, contains intermediate-traffic snapshots related to working days, i.e., 8 :00 to 9 :00 and 20 :00 to 22 :00, as well as week-end daytimes. Class C0 includes low-activity snapshots related to night and early morning hours, i.e., 0 :00 to 8 :00 and 22 :00 to 24 :00.

5.2.2 Call-in profile classification

We depict in Fig. **??** the classification for all the snapshots of the 2-month dataset with measures $\mathcal{V}$ and $\mathcal{D}$.



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

FIGURE 8: Milan. Top-10 score versus the number of clusters (i.e., the dendrogram aggregation level) for the median week training set.

The $\mathcal{D}$ measure identifies four categories, in Fig. 9b, in terms of traffic distribution. Class C1 mainly includes snapshots related to working time, i.e., Monday to Friday from 8 :00 to 18 :00, and presents the highest concentration of call-in activity in the city center, with a much lower relative traffic in all the other zones. Interestingly, also Saturday and Sunday nights, from 0 :00 to 4 :00, belong to this class, thus suggesting that night life in Milan is mainly concentrated in downtown. Class C2 is characterized instead by a more even traffic distribution in all the zones of the city. From a time perspective, this class includes night to early morning (18 :00 to 3 :00 + 1 day), morning (7 :00 to 9 :00) and most of weekend (9 :00 to 0 :00) hours. Class C0 is related to early morning hours, i.e., 6 :00 to 7 :00 during week-days and 6 :00 to 8 :00 during week-end, with a medium-low traffic concentration in the city center and high demand in Zones 8, 9 and 2. Such zones include important transportation hubs (e.g., metro and train stations) or industrial areas. Finally, class C3 groups deep night hours (3 :00 to 5 :00), featuring a medium-high load in the city center and a high demand in industrial or high-density residential areas of Zones 2 and 8.

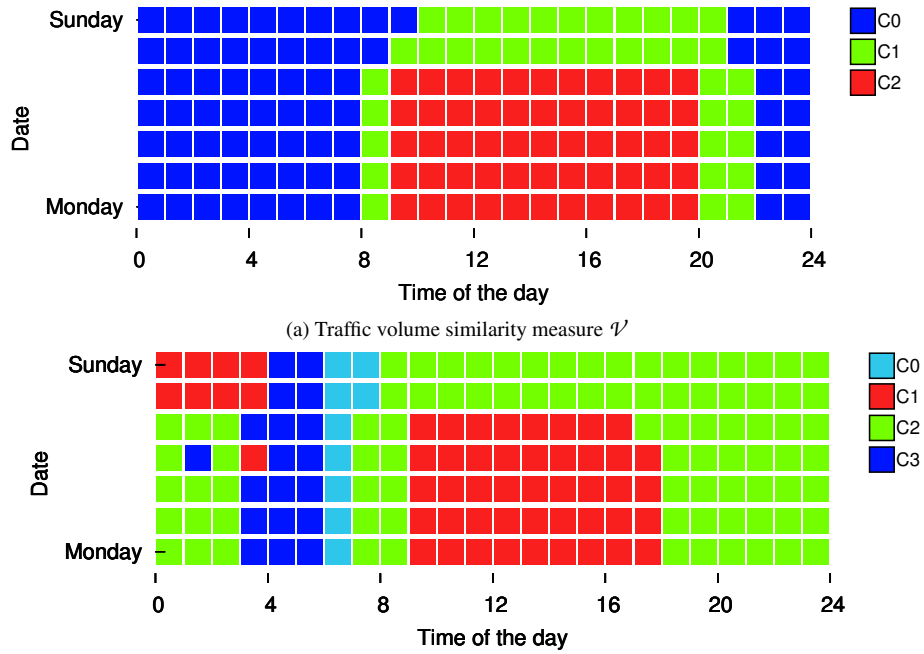
Regarding the $\mathcal{V}$ measure, we remark that most of the snapshots fall in the expected category, with relevant exceptions for the Christmas week (from December 24$^{\text{th}}$ to January 1$^{\text{st}}$) and the All Saints Day public holiday (i.e., November 1$^{\text{st}}$). During Christmas, fewer calls are observed in the whole urban area of Milan especially during working hours, thus highlighting a significant reduced level of business activity, especially in the city center.

Concerning the $\mathcal{D}$ measure, the two major classes, i.e., C1 and C2, are significantly affected by the unusual mobile subscriber behavior during the Christmas week, while they remain almost invariant for the rest of the observation period. Conversely, the two minor clusters, i.e., C0 and C3, which are related to night or first-morning hours and characterized by very low volumes of call-in activity, exhibit much more sensitivity to unpredictable events. Therefore a much higher number of outliers is detected during the corresponding hours.

5.2.3 Call-in profile outliers

As anticipated, the classification performed according to the $\mathcal{V}$ similarity unveils untypical behaviors especially during the Christmas week. As an example, the call-in activity on Christmas and Saint Stephan's afternoons (Dec. 25 and Dec. 26, 14 :00 to 16 :00), in Fig. 11a, is classified in the C0 class instead of C2, because of the much lower call-in volume with respect to typical working days at the same time, Fig. 11b. Similarly, other snapshots on Christmas day, e.g., 10 :00 to 14 :00, and other public holidays, e.g., November 1$^{\text{st}}$ or January 1$^{\text{st}}$, join the intermediate-traffic cluster C1 instead of the expected C2.

Conversely, Wednesday, January 1$^{\text{st}}$ at 0 :00, in Fig. 11c, joins the intermediate-traffic class C1, instead of the expected low-volume class C0 for a typical snapshot at 0 :00, in Fig. 11d. This is due to a very high traffic, mostly concentrated in the first minutes after midnight and quite evenly distributed in the different zones of the city. Similar behaviors can be observed for outgoing calls as well.



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

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

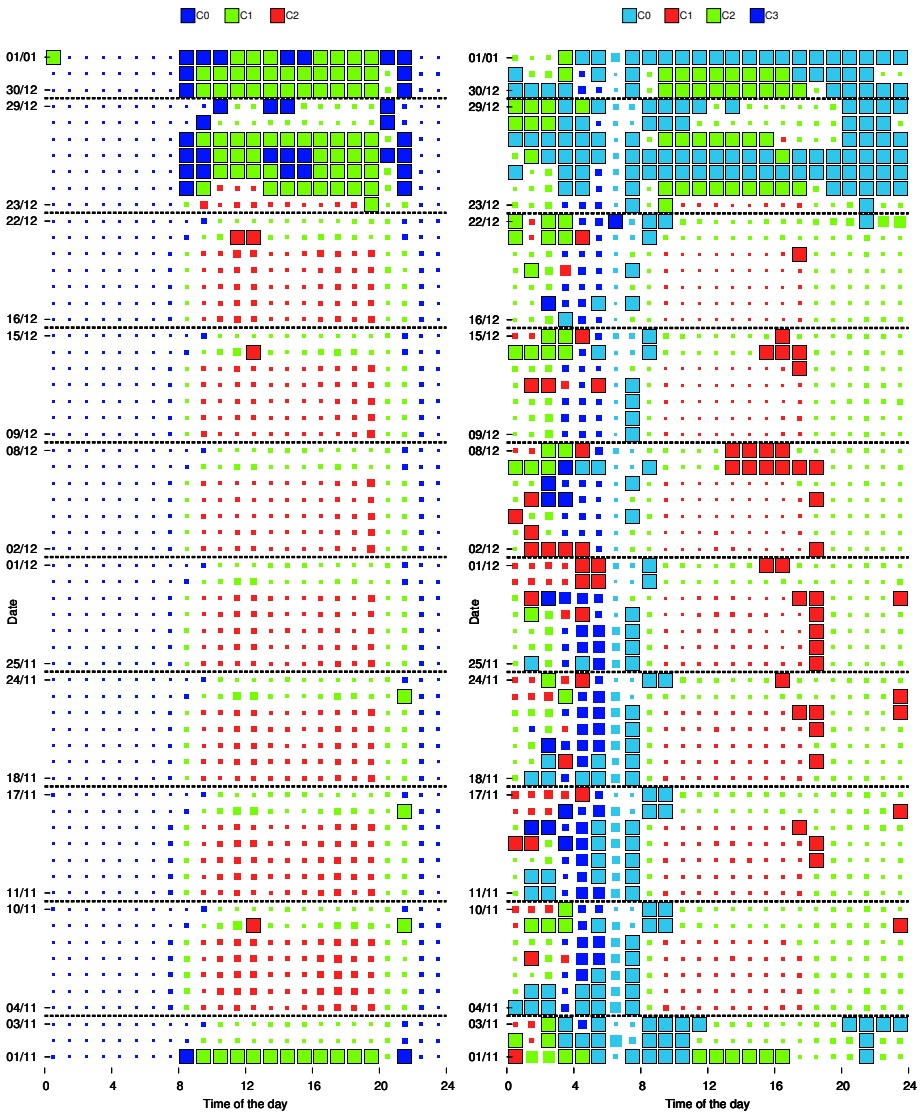
FIGURE 9: Mobile traffic profile categories defined on the training set $\overline{\mathbb{T}}'$. Each square represents one snapshot, and colors map to categories.

Concerning other kinds of communication activities, like SMS-out, January 1st at 0 :00 is classified into the highest-volume class C2, which unveils the habit in Italy of exchanging short text messages for new-year wishes as a preferred option to calls. Short messages continue to appear in higher volumes, with respect to calls, also after midnight, as January 1st at 1 :00 is classified in the C1 cluster for SMS-out, while it remains in C0 for call-in and call-out. Interestingly, the Internet activity at midnight is not affected by significant volume increase, and is classified in the expected C0 even though some peaks are registered in the city center.

Relevant outliers are also detected by the $\mathcal{D}$ measure, as follows. Most of the snapshots related to the Christmas day and other public holidays, e.g., in Fig. 11a, are classified in the C0 cluster, instead of C1. Call-in activity for these snapshots is quite low in the city center, with respect to typical working days, and mostly concentrated in transport and residential areas (e.g., Zones 7, 8 and 2). This is explained by the fact that most of the business activities are closed and the presence of people in the urban area of Milan is significantly reduced with respect to typical days. Interestingly, the framework classifies such snapshots in class C0, together with typical working days at 6 :00, see Fig. 11e. This is due to the fact that, similarly to the holidays-related snapshots above, most of the call-in activity is generated at important transportation hubs (e.g., the central train station in Zone 2, Garibaldi train station in Zone 8 or Piazza Venezia bus stop in zone 7), while a lower traffic is observed in Zone 1, where schools and shops are closed early in the morning.

Some other snapshots, which are very close in time to public holidays (e.g., Tuesday, December 24th or Monday, December 30th, 9 :00 to 18 :00), are classified in the C2 cluster instead of C1. They present a higher activity in the city center and a quite regular traffic distribution in all the other zones and therefore clustered together with evening hours or week-ends, i.e., Fig. 11f.

Like for Abidjan, the framework is capable of detecting special events that occurred in Milan during November and December 2013. For example, on December 7th afternoon, the "La Scala" theater season opening traditionally takes place in the city center of Milan. Such unique event involves the participation by the highest representatives of the Italian State and Italy's most famous economic and cultural personalities. The same day is also the feast day of Saint Ambrose, the patron saint of Milan, with several public and religious events celebrated in the city center. The highest call-in activity on that day is indeed registered



(a) Similarity measure $\mathcal{V}$.

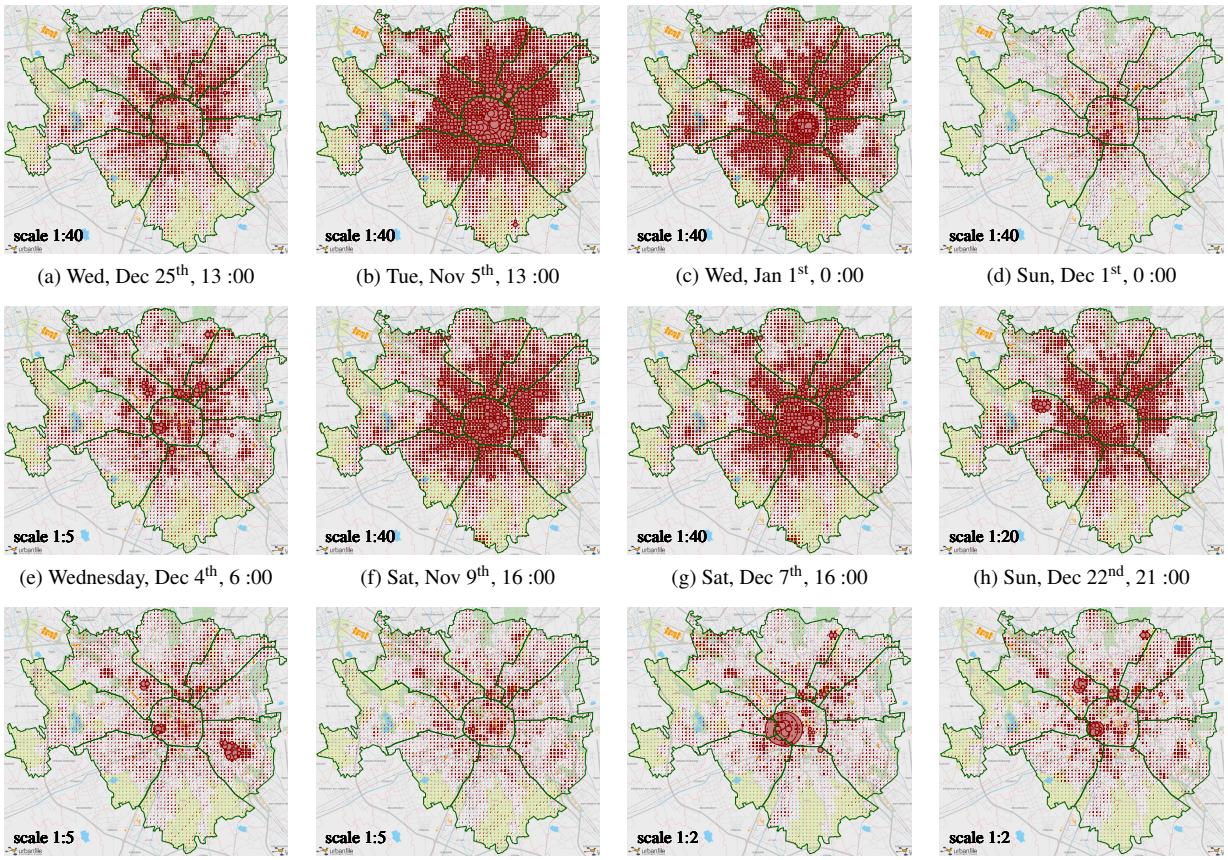
(b) Similarity measure $\mathcal{D}$.

FIGURE 10: Classification of the 2-month Milan traffic data, call-in activity.

at the cells around la Scala theater, in Fig. 11g, and in several other cells in Zone 1. This explains why the framework classifies all the snapshots on December 7$^{\text{th}}$ from 13 :00 to 19 :00 in class C1 instead of C2, together with snapshots related to typical working-time, in Fig. 11b.

Similarly, on Sunday, December 22$^{\text{nd}}$ at 20 :45, one of the most important football matches of the 2014-2015 season, i.e., Inter-Milan, was played in San Siro stadium, registering a record presence of 79,311 spectators. An outlying behavior is detected by our framework, in Fig. 11h, which classifies the snapshot at 21 :00 in class C0, i.e., the one with the lowest concentration in the city center, together with a high traffic in Zone 7 (where the San Siro stadium is located), 8 and 2. The snapshots at 22 :00 and 23 :00 present instead a very high distance from the expected cluster C2, underscoring the untypical call-in activity around the San Siro area.

Another interesting outlying behavior that can be detected by looking at Fig. 10b concerns a set of snapshots in November, related to week-days from 4 :00 to 6 :00, e.g., that in Fig. 11i. Such snapshots are hardly classified in class C3, if compared to the corresponding ones in December. The reason is clear when noting that a relevant amount of call-in activity, generated at cells located around "Mercato Ortofrutticolo" in Zone 4 during November, disappears starting from the first days of December, most probably due to network or



(a) Wed, Dec 25th, 13 :00 (b) Tue, Nov 5th, 13 :00 (c) Wed, Jan 1st, 0 :00 (d) Sun, Dec 1st, 0 :00

(e) Wednesday, Dec 4th, 6 :00 (f) Sat, Nov 9th, 16 :00 (g) Sat, Dec 7th, 16 :00 (h) Sun, Dec 22nd, 21 :00

(i) Fri, Nov 15th, 5 :00 (j) Fri, Dec 13th, 5 :00 (k) Mon, Dec 2nd, 3 :00 (l) Mon, Dec 9th, 3 :00

FIGURE 11: Snapshots of call-in volumes in Milan. Grid squares map to dots, whose size is proportional to the traffic volume in the indicated scale.

collection probe problems, see Fig. 11j.

Finally, some outlying behaviors during night hours, detected by the $\mathcal{D}$ measure, are related to unpredictable and difficult-to-explain events, but are actually confirmed by analysis of the traffic distribution. As an example, the snapshots related to Monday, December 2nd from 1 :00 to 4 :00, in Fig. 11k, are classified in C1, instead of C2 or C3 (see Fig. 11l for a typical Monday at 3 :00), because of an unusual call-in activity at some cells of the city center (very close to a jail and a police station). Such volume increases, which are quite low in absolute value, become significant during night, in the context of an overall very low traffic.

A summary of the most relevant outliers identified in the Milan case study is in Tab. 2. The table presents unusual behaviors in the mobile traffic demand, along with their underlying reason, as detected in different types of traffic data (i.e., call-in, call-out, SMS-in, SMS-out and Internet). Also in the Milan scenario, our framework detected a set of outliers much larger than that we could discuss here or list in Tab. 2. We thus refer the interested Reader to the additional results in App. B and outlier listing in Tab. 2 of the supplemental material.

6 Discussion and open issues

The proposed framework profiles the user demand recorded within mobile networks, as shown by real-world examples using mobile traffic collected in two urban areas. In both case studies, the framework could infer a limited set of significant network usage categories. Each category maps to one specific profile of the network-wide mobile traffic load.

A first interesting observation is that the categories identified in our two reference urban areas are very different. Indeed, each usage profile is representative of the spatiotemporal routines of the subscriber po-

TABLE 2: List of outlying snapshots, according to the classification for Milan with $\mathcal{V}$ and $\mathcal{D}$, respectively (see supplemental material for a more detailed list).

Date	*Similarity measure*	*Activity*	*Actual category*	*Expected category*	*Event*
Wed, Jan 1st, 0 :00	$\mathcal{V}$	Call-in, Call-out, SMS-in	C1	C0	First hour of the new year.
Wed, Jan 1st, 0 :00–1 :00	$\mathcal{V}$	SMS-out	C2	C0	First hour of the new year
Wed, Jan 1st, 1 :00–2 :00	$\mathcal{V}$	SMS-out	C1	C0	Second hour of the new year
Nov 1st, Dec 26th, 30th, 31st 9 :00–20 :00	$\mathcal{V}$	Call-in, Call-out	C1	C2	Public holidays or Christmas-close Days
Nov 1st, Dec 24th, 30th, 31st, 7 :00–8 :00 & 21 :00–22 :00 ; Dec 25th, 26th, 7 :00–9 :00 & 21 :00–22 :00 ; Wed, Jan 1st, 7 :00–11 :00 & 20 :00–22 :00	$\mathcal{V}$	Call-in, Call-out	C0	C1	Public holidays or Christmas-close Days
Fri, Nov 1st, 0 :00	$\mathcal{V}$	SMS-out	C1	C0	Halloween Night
Dec 25th, 26th, Jan 1st, 14 :00–16 :00 ;	$\mathcal{V}$	Call-in, Call-out, SMS-out	C0	C2	Christmas day and Saint Stephan day, lunch time
Fri, Nov 1st, 9 :00–19 :00, Mon, Dec 23rd, 9 :00–19 :00	$\mathcal{V}$	Internet	C1	C2	Public holidays or Christmas-close Days
Dec 25th, 26th, 30th, Jan 1st, 9 :00–20 :00 ;	$\mathcal{V}$	Internet	C0	C2	Public holidays or Christmas-close Days
Fri, Nov 1st, 0 :00	$\mathcal{D}$	Call-in	C1	C2	Halloween night.
Fri, Nov 1st, 3 :00–5 :00	$\mathcal{D}$	Call-in	C2	C3	Halloween night. All Saints day, early morning
Fri, Nov 1st, 7 :00–11 :00	$\mathcal{D}$	Call-in	C0	C2 or C1	All Saints day, morning
Fri, Nov 1st, 11 :00–17 :00	$\mathcal{D}$	Call-in	C2	C1	All Saints day, late morning and afternoon
Fri, Nov 3rd and 4th, 20 :00–3 :00	$\mathcal{D}$	Call-in	C2	C1	Homecoming from All Saints long weekend
Sat, Dec 7th, 13 :00–19 :00	$\mathcal{D}$	Call-in	C1	C2	Saint Ambrose, patron of Milan, and Season opening at "la Scala" Theater at 18 :00
Sat, Dec 14th, 15 :00–18 :00	$\mathcal{D}$	Call-in	C1	C2	Christmas concert in Duomo, performance at La Scala theater
Sun, Dec 22th, 20 :00–0 :00	$\mathcal{D}$	Call-in	C2 (High distance)	C2	Inter-Milan football match, 79,311 spectators in San Siro
Working days, Nov 4 :00–6 :00	$\mathcal{D}$	Call-in	C3 (High distance)	C3	Traffic in "Mercato Ortofrutticolo", missing during December
Dec 25th, 26th and Jan 1st 7 :00–0 :00	$\mathcal{D}$	Call-in	C0	C2 or C1	Christmas holidays, working time
Dec 24th, 27th, 30th and 31st 9 :00–16 :00	$\mathcal{D}$	Call-in	C2	C1	Holidays-close days, working time
Nov 1st, Dec 24th–26th, 30th, 31st, Jan 1st 9 :00 –18 :00	$\mathcal{D}$	Internet	C0	C1	Holidays (or close to holidays) working time
Sat, Nov 23th, 0 :00–3 :00	$\mathcal{D}$	Call-out	C1	C6 or C10	High activity in Navigli area

pulation. Thus, our framework unveils how Abidjan and Milan have dissimilar *pulses* over typical weekly periods. On the one hand, this difference could be expected, due to the different nature of the two cities, similar in size but located in developing and developed countries, respectively. Yet, our analysis unveils exactly when and where differences emerge. On the other hand, this same diversity demonstrates the need for a dedicated profiling for urban regions, since cities may display unique habits that result in singular spatiotemporal fluctuations of the mobile demand.

The latter point introduces the second element of discussion, i.e., the fact that understanding the exact demand fluctuations in each urban region is paramount to a successful deployment of cellular cognitive networks. Indeed, the presence of such a strong variability in the way the mobile network is used underscores the inadequacy of static resource allocations. Also, it highlights how the replication of identical policies for dynamic resource management may not fit different scenarios, characterized by dissimilar spatiotemporal pulses of the population activity.

In this context, our framework can be leveraged to determine a first macroscopic separation among the distinguishable profiles of the mobile demand in a target region. As it is unsupervised and operates in a completely automated fashion, the framework can provide input to cellular cognitive network management functions, and thus support an adaptive release and relocation of resources that is tailored to the actual demand. In addition, the capability of detecting unusual network usages can be exploited to adapt the resource allocation on-the-fly, at least for events occurring on timescales in the order of a few hours. Possible cognitive networking applications that could take benefit from this profiling function are C-RAN, base station switch on/off, or accommodating M2M communication. However, we stress that this work just represents a preliminary step to the development of mobile traffic analytics for future cognitive mobile networks. As such, it has limitations that open future research directions.

Specifically, in both our case studies and under any combination of similarity measure and traffic type, the profiles obtained by our unsupervised framework only outline macroscopic variations of the mobile demand. Indeed, the number of categories is small, capturing large variations of network usage, in space (i.e., fluctuations of the demand across urban zones spanning several km^2 each) and time (i.e., differences emerging among groups of several hours each). This is a quite natural outcome, since the framework targets the separation of mobile demand profiles that maximizes their diversity, and the most evident differences emerge at such macroscopic scales. However, a higher resolution may be required for a fine-grained resource allocation : in this case, the approach presented in this paper shall mainly provide a first filtering upon which to run further refinements. Several extensions to our work are possible in this direction. One simple option is intersecting the categories obtained with different similarity measures into a larger number of more specialized categories. Another option is investigating the local maxima of the top-*k* index that often appear in presence of a high number of clusters. A third option is recursively running the framework within each profile separately, disaggregating them into increasingly precise categories. Overall, the study of more fine-grained variations in the mobile traffic demand is an aspect that we intend to investigate in our future works.

Another item for future work is the evaluation of a much larger set of similarity measures and traffic types. As a matter of fact, the similarity metrics we explored in this work are basic ones, and more complex measures can be envisioned. They could lead to different profiles, whose correlation with those identified by the present version of the framework needs to be assessed. Along an orthogonal direction, descending at the application level, and profiling the dynamics of the demand on a per-service basis, would offer valuable information for resource management, since network requirements depend on the nature of the service.

Finally, our results clearly show that different scenarios can yield very diverse profiles of the mobile demand. It is thus important to evaluate the framework on a significant number of heterogeneous case studies, so as to generalize the conclusions on its performance. At the same time, an extensive evaluation campaign would allow unveiling the spatiotemporal variability that characterizes mobile network usages in a substantial set of situations, possibly identifying correlations in the profiles emerging in regions that share topological, social, economical, or political features. In turn, such a study shall pinpoint the external factors that affect the most the mobile demand dynamics. The same consideration holds for different kinds of mobile networks, based on, e.g., Wi-Fi or femtocell deployments : indeed, our proposed methodology can be readily applied to traffic data recorded at networks that rely on different radio access technologies.

7 Acknowledgments

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. IEEE Communications Magazine, 47(4) :52-59, Apr. 2009.

Références

[1] D. Naboulsi, R. Stanica, and M. Fiore, “Classifying Call Profiles in Large-Scale Mobile Traffic Datasets,” *Proc. IEEE INFOCOM 2014*, Toronto, ON, Canada, Apr. 2014.

[2] “Cisco visual networking index : Global mobile data traffic forecast update, 2013-2018,” Feb 2014.

[3] Key technological challenges of the EC H2020 5G Infrastructure PPP. [Online]. Available : http://5g-ppp.eu/

[4] “EC H2020 5G Infrastructure PPP. Pre-structuring Model, version 2.0.” April 2014.

[5] R. Thomas, L. DaSilva, and A. MacKenzie, “Cognitive Networks,” *Proc. IEEE DySPAN 2005*, Baltimore, MD, USA, Nov. 2005.

[6] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, “Large-scale Mobile Traffic Analysis : A Survey,” *hal preprint hal-01132385*, 2015.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding Individual Human Mobility Patterns,” *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[8] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human Mobility Modeling at Metropolitan Scales," *Proc. ACM MobiSys 2012*, Low Wood Bay, UK, Jun. 2012.

[9] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[10] R. A. Becker, R. Caceres, K. Hanson, J. M. 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, 2011.

[11] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven Mobile Data Traffic Modeling in a Large Metropolitan Area," *Proc. IEEE PerCom 2015*, St. Louis, MO, USA, Mar. 2015.

[12] 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," *Proc. ACM UrbaNe 2012*, Nice, France, Dec. 2012.

[13] D. Goergen, V. Mendiratta, R. State, and T. Engel, "Identifying Abnormal Patterns in Cellular Communication Flows," *Proc. ACM IPTComm 2013*, Chicago, IL, USA, Oct. 2013.

[14] Orange D4D Challenge. [Online]. http://www.d4d.orange.com

[15] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology," *Proc. ACM MobiHoc 2015*, Hangzhou, China, Jun. 2015.

[16] Telecom Italia Big Data Challenge. [Online]. Available : http://www.telecomitalia.com/tit/en/bigdatachallenge.html

[17] 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.

[18] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering Individual and Collective Human Dynamics from Mobile Phone Records," *Journal of Physics A : Mathematical and Theoretical*, vol. 41, no. 22, p. 224015, 2008.

[19] P. Paraskevopoulos, T.-C. Dinh, Z. Dashdorj, T. Palpanas, and L. Serafini, "Identification and Characterization of Human Behavior Patterns from Mobile Phone Data," *Proc. NetMob 2013*, Cambridge, MA, USA, Apr. 2013.

[20] R. Pulselli, P. Ramono, C. Ratti, and E. Tiezzi, "Computing Urban Mobile Landscapes through Monitoring Population Density based on Cellphone Chatting," *International Journal of Design and Nature and Ecodynamics*, vol. 3, no. 2, pp. 121–134, 2008.

[21] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, "Towards Estimating the Presence of Visitors from the Aggregate Mobile Phone Network Activity They Generate," *Proc. CUPUM 2009*, Hong Kong, China, Jun. 2009.

[22] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A First Look at Cellular Network Performance during Crowded Events," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 1, pp. 17–28, 2013.

[23] F. Yu, G. Xue, H. Zhu, Z. Hu, M. Li, and G. Zhang, "Cutting without Pain : Mitigating 3G Radio Tail Effect on Smartphones," *Proc. IEEE INFOCOM 2013*, Turin, Italy, Apr. 2013.

[24] A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and J. Erman, "To Cache or Not to Cache : The 3G Case," *IEEE Internet Computing*, vol. 15, no. 2, pp. 27–34, 2011.

[25] Y. Zhu, C. Zhang, and Y. Wang, "Mobile Data Delivery through Opportunistic Communications among Cellular Users : A Case Study for the D4D Challenge", *Proc. NetMob 2013*, Cambridge, MA, USA, Apr. 2013.

[26] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G.-K. Chang, "The Case for Re-configurable Backhaul in Cloud-RAN based Small Cell Networks," *Proc. IEEE INFOCOM 2013*, Turin, Italy, Apr. 2013.

[27] E. Oh, K. Son, and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp.2126–2136, 2013.

[28] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio Resource Allocation in LTE-Advanced Cellular Networks with M2M Communications," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 184–192, 2012.

[29] R. Xu, D. Wunsch *et al.*, "Survey of Clustering Algorithms," *IEEE Transaction on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[30] G. W. Milligan and M. C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[31] T. Caliński and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[32] E. Beale, *Euclidean Cluster Analysis*, Scientific Control Systems Limited, 1969.

[33] L. J. Hubert and J. R. Levin, "A General Statistical Framework for Assessing Categorical Clustering in Free Recall," *Psychological Bulletin*, vol. 83, no. 6, p. 1072, 1976.

[34] R. O. Duda, and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, 1973.

[35] P. J. Rousseeuw, "Silhouettes : A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[36] J. A. Hartigan, *Clustering Algorithms*, John Wiley and Sons, 1975.

[37] W. J. Krzanowski and Y. Lai, "A Criterion for Determining the Number of Groups in a Data Set using Sum-of-Squares Clustering," *Biometrics*, vol. 44, no. 1, pp. 23–34, 1988.

[38] R. Guigourès, D. Gay, M. Boullé, F. Clérot, and F. Rossi, "Country-scale Exploratory Analysis of Call Detail Records through the Lens of Data Grid Models," *arXiv preprint arXiv :1503.06060*, 2015.