DOCTORAL SCHOOL InfoMaths

PHD THESIS

to obtain the title of

Doctor of Philosophy

of INSA Lyon

Specialty : Computer Science

Prepared at the CITI Lab, in theUrbanet Team

Analysis and exploitation of mobile traffic datasets

Defended by

Diala Naboulsi

on September 24, 2015

Jury :

President :	Eric Fleury	-	Professeur des Universités ENS de Lyon, France
Reviewers :	Roch Glitho	-	Associate Professor Concordia University, Canada
	Xavier Lagrange	-	Professeur des Universités Télécom Bretagne, France
	Rami Langar	-	Maitre de Conférences, HDR UPMC, France
Examinator :	Thierry Turletti	-	Directeur de Recherche Inria Sophia Antoplis, France
Advisors :	Marco Fiore	-	Chercheur, HDR CNR - IEIIT, Italie
	Razvan Stanica	-	Maitre de Conférences INSA de Lyon, France
Director :	Fabrice Valois	-	Professeur des Universités INSA de Lyon, France

“The teacher who is indeed wise does not bid you to enter the house of his wisdom but rather leads you to the threshold of your mind.”

Khalil Gibran

Remerciements

Tout d'abord, je tiens à remercier M. Fabrice Valois de m'avoir donné la possibilité de poursuivre ma thèse. Je le remercie pour tous ses conseils constructifs, nos discussions enrichissantes et son support aux moments les plus stressants de la vie de doctorante.

Je remercie M. Marco Fiore, d'avoir été toujours prêt à partager généreusement son expérience et son expertise tout au long de ces trois années. Marco m'a appris beaucoup de notions sur la recherche et j'en serai pour toujours reconnaissante.

Je remercie M. Razvan Stanica pour son altruisme, sa modestie, sa brillance, et son encadrement extrêmement pédagogique et amusant à la fois. Merci d'avoir été toujours disponible pour m'écouter, suivre mes idées jusqu'au bout et m'aider à les améliorer. Travailler avec Razvan, ce n'est que du plaisir.

J'adresse mes remerciements aussi à tous les membres du jury pour avoir accepté d'évaluer les travaux menés dans le cadre de cette thèse.

Je remercie M. Hervé Rivano et tous les membres de l'équipe UrbaNet pour nos discussions intéressantes et les moments agréables que nous avons passés ensemble. Je remercie Trista, ma chère amie et collègue de bureau pour toutes les blagues et recettes que nous avons échangées. Je remercie aussi Gaelle pour son aide avec les procédures administratives compliquées et pour son grand sourire.

Je remercie les membres du laboratoire CITI, du département TC à l'INSA de Lyon, et du département informatique à l'Université Lyon 1 pour les ambiances de travail amicales et professionnelles.

Je remercie tous mes amis en France pour les moments inoubliables que nous avons partagés. Je remercie également tous mes amis au Liban, pour leurs encouragements constants.

Je remercie toute ma famille, mes parents, ma soeur et mon frère, pour tous les efforts qu'ils ont mené pour que je puisse commencer et poursuivre ma thèse. Leur support est un pilier fondateur de tout ce que j'ai fait.

Enfin, je remercie Roger pour sa patience, son soutien et son amour inconditionnels.

Contents

List of Figures

List of Tables

Part I

Cellular Networks and Mobile Traffic Datasets: an Introduction

Chapter 1

Introduction

1.1 Context

From their early conception, cellular networks have played a major role in progressively shaping the digital dimension of our societies. Starting from their first generation (1G), in the 80s, cellular networks have enabled users with mobile communication capabilities by providing analog mobile voice services. Later, the second generation (2G) networks, as of the early 90s, introduced digital voice and texting services. Additionally, they paved the way towards data communications, which were then framed as part of the third generation (3G) networks, at the beginning of our millennium. While 3G networks provide data rates of a few Mb/s, their successors, fourth generation networks (4G), pushed this limit to hundreds of Mb/s.

The evolution of mobile subscriptions continues to testify the success of cellular communication systems with a global mobile penetration rate that surpassed 94% in 2014 [1]. However, this massive integration of mobile devices in our lives is translating into an explosion of global mobile data traffic. The latter has witnessed an important increase over the last few years reaching 2.5 ExaBytes per month in 2014 [1, 2]. Moreover, with respect to 2014, it is expected to grow ten-fold by 2019 [2].

Consequently, mobile network operators are currently facing significant challenges. On the economic plan, their revenues are not growing at the same pace of the traffic load, making further investments, targeting the improvement of the network, harder to consider. On the technical plan, challenges span from increasing the network capacity, and providing higher data rates, to designing energy efficient solutions. Additionally, the diversity of mobile devices, with heterogeneous traffic patterns, is introducing another layer of complexity. As an example, mobile phones and Machine-to-Machine (M2M) communications display quite different traffic patterns [4] with distinct requirements that operators would need to take into account for an optimal system performance.

All this has triggered research efforts by academic and industrial institutions, who are actively seeking for networking solutions, targeting the conception, design and management of future fifth generation (5G) networks. Adequate solutions are envisioned to considerably improve the overall performance of the network at lowest possible costs.

However, while the networking community has largely focused on the problems that would result from the proliferation of mobile devices, it has not considered the potential they hold. In reality, these devices constantly interact with the network infrastructure and their activity is recorded by network operators, typically for monitoring and billing purposes. The resulting logs, to which we refer as mobile traffic datasets, convey important information concerning spatio-temporal traffic dynamics, relating to large sets of equipments that can include thousands and millions of devices.

When it comes to mobile phone logs, such datasets carry promising potential for a variety of fields [5, 6] including, among others, sociology [7, 8], mobility [9, 10], urban planning [11] and networking [12, 13]. The reason is that they allow to track individual and aggregate human activities over time and space at unprecedented scales and at relatively no cost.

For the field of cellular networks, mobile traffic datasets constitute a very useful component. They can bring mobile networking research efforts a step closer to real-world systems, at a time in which it takes a decade to drive networking solutions, from their early research stage, to their actual implementations, in industry.

This is especially relevant for traffic-oriented network management solutions that adapt to the evolution of traffic over space and time. On the one hand, analyzing mobile traffic datasets would allow to understand how traffic characteristics change over these two dimensions. This can help devise design guidelines for such solutions. On the other hand, mobile traffic datasets constitute very useful tools for the evaluation of network solutions. Clearly, in order to capture an accurate view of their performance, in an actual network, large-scale realistic evaluation environments are required. In contrast to expensive real-world experiments, mobile traffic datasets can play this role, at relatively no cost.

The work presented in this thesis sheds light on the potential of mobile traffic datasets, with respect to these two aspects. We review its main contributions in Sec. 1.2 and its organization in Sec. 1.3.

1.2 Thesis contributions

The thesis covers the main following contributions:

- **Identification of key findings of previous mobile traffic datasets analyses and exploitations.** We synthesize principal features of traffic evolution obtained by previous works analyzing mobile traffic datasets, in Chapter 3. These include major traffic characteristics over space and time, from an aggregate perspective, i.e. relating to all users accessing the network, and an individual perspective, i.e. relating to the behavior of each device by itself. In addition to that, we shed light on the few exploitations of mobile traffic datasets relevant for cellular networks in Chapter 5. We stress the fact that these studies mostly consider static strategies, while mobile traffic datasets carry promising potential for more beneficial dynamic network solutions.

- **Extraction of representative network-wide mobile traffic utilization patterns.** We propose a mobile traffic datasets characterization framework, capable of processing very large datasets and automatically outlining network-wide utilization patterns, in Chapter 4. The framework combines a set of data mining tools and builds categories of mobile traffic usages by grouping together similar patterns according to a couple of distance measures that can capture complementary facets of mobile traffic profiles. By applying the framework over a real-world traffic dataset, we unveil its capacity to generate meaningful utilization patterns.

- **Detection of outlying mobile traffic usage behaviors.** As part of the mobile traffic datasets characterization framework, proposed in Chapter 4, we perform a classification step that allows us to separate between typical and outlying mobile traffic behaviors. The application of the framework over a real-world traffic dataset shows how it allows to identify unexpected behaviors in the mobile demand, which we are able to map to underlying real-world events.

- **Evaluation of achievable energy savings over cellular networks infrastructure.** We introduce a methodology that allows to reduce energy consumption in cellular networks, based on a power control mechanism, at the level of individual base stations, in Chapter 6. We show that the proposed strategy adapts network infrastructure energy consumption to the traffic evolution, throughout the day, and can lead to important energy savings.

- **Conception of a dynamic Cloud-Radio Access Network topology configuration method.** We propose a dynamic methodology for the management of future Cloud-Radio Access Networks (C-RAN) topology in Chapter 7. Our method allows us to underscore the potential of a C-RAN architecture for the management of users mobility, with respect to a traditional decentralized access network. By applying our scheme over real-world traffic datasets, for two mobility scenarios, we show that significant reductions in terms of handovers can be obtained. Additionally, our results suggest promising energy and computational efforts savings over the network infrastructure.

1.3 Thesis organization

The thesis is divided into eight chapters, forming four separate parts as follows.

The first part of the thesis, *Cellular Networks and Mobile Traffic Datasets: an Introduction*, includes the current chapter, as well as Chapter 2. It aims at introducing the context of the work and presenting general aspects relating to the topic of the thesis. In Chapter 2, we cover global features of mobile traffic datasets. We start by presenting an overview of cellular network architectures, allowing us to describe mobile traffic datasets collection operation, through different monitoring probes. We then discuss general characteristics of mobile traffic datasets, followed by a detailed presentation of the mobile traffic dataset that we employ throughout the thesis. Additionally, we explain how we process the dataset in order to employ it for networking studies.

The second part of the manuscript, *A Networking Perspective on Mobile Traffic Datasets Analysis*, and the third part, *Networking Solutions and Mobile Traffic Datasets*, are dedicated to the discussion of our contributions with complementary facets. The second part, organized into two chapters, Chapter 3 and 4, targets networking analysis of mobile traffic datasets.

We focus on previous works analyzing mobile traffic datasets, in Chapter 3. We begin by introducing a set of tools and methods, widely used for studying the evolution of mobile traffic. We then cover works that analyze mobile traffic datasets from an aggregate point of view, accounting for the overall traffic properties, with respect to all users accessing the network. After that, we present findings of previous studies characterizing mobile traffic, from an individual point of view, and thus considering the behavior of each device by itself.

Chapter 4 is dedicated to the presentation and application of our mobile traffic datasets characterization framework. There, we detail the different components of the framework. We present the system model that we employ together with distance measures capable of capturing complementary sides of mobile traffic datasets, for the comparison of usage patterns. Then, we explain how we define usage profile categories that we employ in a following classification step to tell apart typical and outlying behaviors. In parallel to that, we apply the framework over real-world traffic datasets that testify its capabilities.

In turn, the third part, divided into three chapters, Chapter 5, 6, and 7, sheds light on the network-oriented exploitation of mobile traffic datasets. Chapter 5 provides an overview of previous studies exploiting mobile traffic datasets for innovative solutions relating to cellular networks. We discuss these studies covering separately marketing strategies, and technical solutions.

In Chapter 6, we focus on the application of mobile traffic datasets for the evaluation of achievable energy savings on the cellular network infrastructure, based on a power control mechanism over individual base stations. We review previous works relating to energy consumption reduction based on such power control scheme and highlight weak aspects of these studies. After that, we propose a strategy that allows to adapt the access network power configuration to the system requirements in terms of geographical coverage and users demand. In addition to that, we employ the strategy over realistic data derived from the real-world traffic datasets, highlighting the capabilities of our strategy.

Chapter 7 targets another networking application, relating to the management of users mobility in future C-RAN networks. Accordingly, we begin by providing an overview of general aspects relevant to the concept of C-RAN. We then review previous works dealing with dynamic C-RAN topology configuration solutions, and position our study with respect to them. Next, we present our dynamic system model and methodology to manage the C-RAN topology. Finally, we unveil the potential of the proposed methodology, by applying it over realistic data.

The last part of the thesis, *Cellular Networks and Mobile Traffic Datasets: Conclusions and Perspectives*, includes Chapter 8, where we report conclusions that we draw from the thesis and discuss possible directions that can be considered in the future.

Chapter 2

Mobile traffic datasets

2.1 Mobile traffic datasets

Mobile traffic datasets provide a description of spatio-temporal usages of customers over the cellular network. These datasets can be collected on the user's side, via dedicated applications that run over the user equipment. Well-known examples include the Lausanne Data Collection Campaign, led by Nokia Research Center as part of the Nokia Mobile Data Challenge [14]. Such datasets are mostly complemented by per-user social profile descriptions and can thus be useful for studies aiming at relating usage consumption to individual lifestyles and habits. However, even with such precise information, the generalization of obtained results remains questionable, as the campaigns are typically limited to a small group of individuals with relatively homogeneous profiles, mostly working in academic institutions.

Richer datasets, on which we focus in this thesis, are those obtained on the cellular network operator's side. Originally, they have been collected by network operators for billing and monitoring purposes. Due to the sensitive information that they hold, network operators have been very cautious about sharing them with other parties. By applying anonymization and information aggregation schemes, allowing to comply with privacy preserving regulations, operators have started sharing such datasets for research and development purposes. The past couple of years have witnessed significant initiatives by two major multinational operators in the form of innovation challenges: the Data for Development (D4D) challenges by Orange [15] and the Telecom Italia Big Data Challenges [16]. Emerging works using these datasets cover a wide spectrum of domains, including sociology [7, 8], epidemiology [17], mobility[9, 10], and networking [12, 13]. In this chapter, we pay particular attention to networking studies relevant to our work, and we refer the reader to [5, 6] for reviews of mobile traffic analyses in other fields.

The rest of this section highlights basic aspects of the collection process of mobile traffic datasets on the operator side. We introduce cellular networks architecture and monitoring probes used to collect the data, and provide a description of their general characteristics. We remark that, in the rest of the document, we employ the expression *mobile traffic datasets* to refer to the operator side collected data.

2.1.1 Architecture of cellular networks

Cellular networks have been designed based on a modular architecture that allows interoperability among different generations. From a functional point of view, the global structure of cellular networks has not changed much and physical entities remain grouped into two domains: Radio Access Network (RAN) and Core Network (CN) domains. The RAN domain provides users with radio resources to access the CN domain, while the latter is responsible for the management of services, including the establishment, termination and reconfiguration of current communications. Fig. 2.1 outlines these domains across 2G, 3G, and 4G networks according to the corresponding major standards, i.e. Global System for Mobile communications (GSM), Universal Mobile Telecommunications System (UMTS) and Long-Term Evolution (LTE) respectively. We remark that on top of the core network, service networks operate in order to enable value-added services and applications. However, our discussion mainly focuses on the physical system architecture, so as to highlight probing operations of interest for the thesis work.

2G GSM Networks: The RAN domain in 2G networks is referred to as Base Station Subsystem (BSS). It is formed by Base Transceiving Stations (BTS) and Base Station Controllers (BSC). A BTS is responsible for radio transmissions and receptions and some physical layer processing, while a BSC controlling a group of BTSs is responsible for the management of radio resources, paging and some handover procedures. The CN domain referred to as Network and Switching Sub-System (NSS), and providing only circuit-switched (CS) services, is instead formed by Mobile Switching Centers (MSC) and Gateway Mobile Switching Centers (GMSC) responsible for voice call management, managing user equipment (UE) registrations and mobility. Several major databases useful for managing customers are also present in the CN domain: the Home Location Register (HLR) and the Visitor Location Register (VLR), typically allowing to locate users, the Authentication Center (AuC), dealing with security-related data for the authentication and encryption procedures, and the Equipment Identity Register (EIR), including data related to mobile equipments.

3G UMTS Networks: The Universal Terrestrial Radio Access Network (UTRAN) represents the RAN in UMTS networks. It is composed of NodeBs and Radio Network Controllers (RNC) playing similar roles to those by BTSs and BSCs in GSM networks. The CN, on the other hand, is divided into two parts: the circuit-switched (CS) and packet-switched (PS) domains, representing practically a combination of the GSM NSS, and the General Packet Radio Service (GPRS) backbone. We remark that GPRS is a technology between 2G and 3G cellular networks,

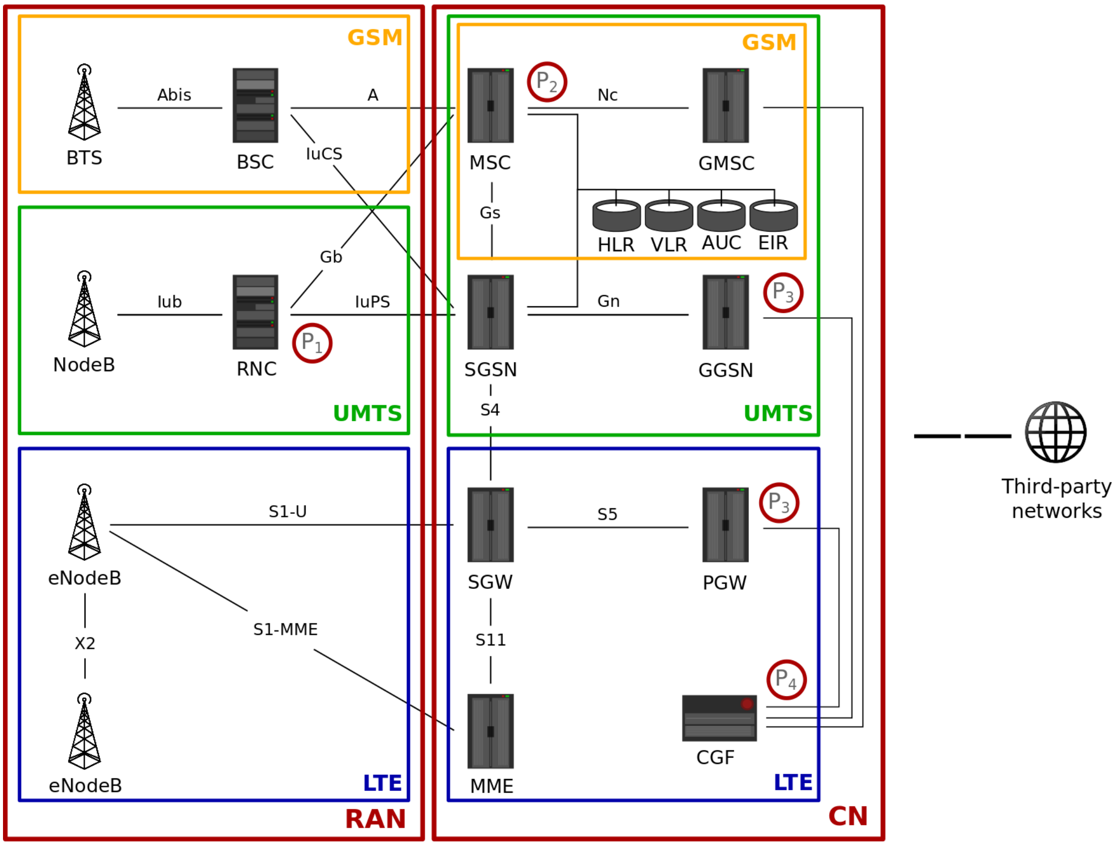


FIGURE 2.1: GSM, UMTS, and LTE cellular networks architectures and mobile traffic monitoring probes.

that provides mobile data services with data rates of a few Kb/s. The PS domain is formed of Serving GPRS Support Node (SGSN) and Gateway GPRS support node (GGSN), responsible for handling packet connections of UEs, security functionalities and mobility management functions as well as data routing.

4G LTE Networks: In contrast to UMTS systems, LTE networks are designed to provide only PS services. The RAN, or Enhanced-UTRAN (EUTRAN) in 4G terminology, is only formed by interconnected base stations (BS) called eNodeBs, with no centralized controlling devices, as opposed to its preceding technologies, but remains responsible for radio-related functions. ENodeBs are rather directly connected to the core network, referred to as Enhanced Packet Core (EPC). The EPC, responsible for the overall control of UEs, includes Serving Gateways (SGW) and Packet Gateways (PGW) managing data packets routing and forwarding, as well as network address allocations, and Mobility Management Entities (MME) performing connection management. Additionally, by cooperating with the following other entities: Home Subscriber Server (HSS), Enhanced Serving Mobile Location Center (E-SMLC), and Gateway Mobile Location Center (GMLC), the MME completes mobility-related and authentication-related tasks. Finally, the EPC also includes a Policy Control and Charging Rules Function (PCRF) entity orchestrating policy and flow control decision makings.

For billing and inter-operator accounting procedures, a set of logical charging function are implemented in the network. They are responsible for collecting network resource usages by each customer. The main functions are the following: the Charging Trigger Function (CTF), which generates charging events based on the observation of network resource usages; the Charging Data Function (CDF), which receives charging events from the CTF to construct Call Detail Records (CDR), providing for each user reports concerning his communications; and the Charging Gateway Function (CGF), responsible for validating, reformatting and storing CDRs before sending them to the billing domain.

2.1.2 Mobile traffic datasets collection

Mobile traffic datasets are obtained via monitoring probes that can be placed over suitable interfaces within the cellular network as indicated in Fig. 2.1. Depending on their position, these probes allow to capture traffic consumption with various levels of granularity as explained next.

RNC probes, marked with 'P_1' in Fig. 2.1, are located at the Iub interface between a NodeB and an RNC in UMTS systems. They are capable of capturing signaling messages related to detailed radio resource management events. They thus allow to detect network attach and detach operations, initiation and termination of calls and sessions, as well as fine-grained and coarse-grained location update events. These probes also provide some performance indicators on data transmission, e.g. uplink and downlink throughput experienced by the user.

MSC probes, marked with 'P_2' in Fig. 2.1, are placed in the CS core network of GSM and UMTS systems, over the N_c interface. They retrieve signalling messages only related to voice and texting services that overcome the BSC and the RNC. Thus, all events locally handled by the latter entities such as intra-BSC or intra-RNC handovers are transparent to these probes.

GGSN/PGW probes, marked with 'P_3' in Fig. 2.1, are located at the border between the PS core and third-party networks, in UMTS and LTE systems. They allow to collect per-session Packet Data Protocol (PDP) Context information. For each user, they can capture the timespan of each session, the resulting traffic volume, and the implied type of service. These probes are not capable of capturing voice and texting activities.

CGF probes, marked with 'P_4' in Fig. 2.1, collect per-user CDR information from the CGF. CDRs typically include for each communication, its start and end times, together with originating cell. In addition, for packet traffic, they contain the quantity of transferred data.

2.1.3 Mobile traffic datasets characteristics

Mobile traffic datasets are very diverse in terms of spatio-temporal granularity, aggregation level and scale. This is a direct consequence of the location of collection probes within the cellular network architecture, privacy-preserving measures, and operator's objective in case the data is publicly shared. The variety of these datasets makes some of them more appealing for researchers than others. In the following, we highlight their major features.

Spatio-temporal granularity: The level of granularity of the collected data is imposed by the collection probe. In general, the closer the probe is to the user, the more detailed is the information it can obtain. RNC probes provide information with the highest level of detail, as they relate users signaling events to their cell location. However, they imply significant deployment and maintenance costs in order to cover network-wide operations. Thus, they are only practical for the operator for some localized monitoring activities, spanning over a small geographical area.

Instead, core network probes are easier to manage, since the corresponding network equipments are typically placed in main central offices. However, these probes, at the exception of the CGF probes, associate recorded events to coarse-grained location information. More critical is the fact that, in some situations, location information is outdated. As an example, this is the case of a PDP context, to which is attributed only the initial user location at the setup of the PDP session, and remains unchanged even in case the user changes his location.

In case of sharing mobile traffic data with other parties, additional procedures can be considered to blur some information for privacy concerns. In terms of network infrastructure, operators consider the position of their equipments as sensitive information. Consequently, unless forced by laws and regulations, they prefer to provide an approximation of their locations, instead of exact ones. As an example, Orange applied such a technique, when sharing the positions of its base stations, in Ivory Coast and Senegal, for the D4D challenges [15]. From the perspective of users, mobile traffic data clearly holds sensitive information relating to their daily life. Applying anonymization schemes that hide users identifiers may not be enough to protect users privacy as discussed in more detail in Sec. 5.3.5. As a result, when sharing per-user data, operators tend to provide coarse-grained spatio-temporal location information. While such schemes can be convenient to protect users data, they limit their usefulness for some fields of study.

Aggregation level: Mobile traffic datasets can present information separated for each user, or aggregated for a group of users. Operators can consider network-oriented aggregation schemes, in order to avoid possible privacy problems resulting from sharing per-user information, while still maintaining high spatio-temporal granularities. As an example, network operators can gather all users communication data for each cell on a periodic temporal basis. Such a scheme

date_time	caller_id	callee_id	call_duration	antenna_code
2012-03-21 15:45:00	130876	156762	45	345

TABLE 2.1: Standard Call Detail Records (CDRs) format.

can still be useful for some studies, such as networking analysis aiming at characterizing usage patterns over the network [19, 20].

Scale: The large-scale capability of mobile traffic data is one of the major reasons that oriented researchers attention towards them, regarding the geographical area they compass, their timespan, and the number of users they can cover. In fact, the datasets can span up to a country-wide level, with information over several months about millions of users. Clearly, for researchers, the larger the scale, the more useful and representative the data. Nevertheless, operators remain cautious about the scale of publicly shared data, again for privacy concerns.

Whether considering the spatio-temporal granularity, the level of aggregation or the scale of data, there is no general rule that determines the optimal measures to ensure users privacy. Finding strategies that balance between the usefulness of data and their granularity and scales remains an open question, in the research community.

2.2 Ivory Coast D4D datasets

In 2013, Orange launched its first Data for Development (D4D) challenge, an innovation challenge aiming at employing mobile traffic datasets towards the development of Ivory Coast [18]. Participants were provided with technical and traffic datasets that reflect users activities over the network. The mobile traffic datasets were derived from a pre-constructed CDR database with information about Orange customers. In the following, we provide more details concerning the different datasets.

2.2.1 Orange CDR database

From December 5th, 2011 until April 22nd, 2012, Orange collected, in a CDR database, information about communication activities of its five million customers in Ivory Coast. These CDRs provide the position of a user, approximated as the BS location, at every time he initiates a call or sends an SMS. Their standard format is indicated in Tab. 2.1.

Information relating to customers that subscribed or terminated contracts at Orange during the data collection phase has been omitted, so as to reflect the behavior of the same set of users over the whole observation period. Additionally, the identifiers of all customers have been replaced

antenna_id	longitude	latitude
1	-4.143452	5.342044
2	-3.913602	5.341612
3	-3.967045	5.263331
4	-4.070007	5.451365
5	-3.496235	6.729410

Table 2.2: Extract from antennas positions dataset.

subpref_id	longitude	latitude
1	-3.260397	6.906417
2	-3.632290	6.907771
3	-3.397551	6.426104
4	-3.662953	6.660800
5	-3.440788	6.937723

Table 2.3: Extract from sub-prefectures positions dataset.

with randomly generated numbers, to prevent a direct mapping between their identities and their mobile phone activities.

2.2.2 Technical datasets

Orange provided two technical datasets that allow to map mobile traffic information over the geographical space.

Antennas positions dataset. In this dataset, Orange provides for each of its antennas, its approximate location in terms of longitude and latitude coordinates. Tab. 2.2 shows an extract from this dataset.

Fig. 2.2 shows the position of these antennas over the map of Ivory Coast, together with its 255 sub-prefectures. This dataset is particularly useful for studies employing the aggregate antenna-to-antenna dataset and the individual high spatial resolution traffic dataset, described in Sec. 2.2.3, including information about calling activities of customers at a per-antenna level.

Sub-prefectures positions dataset. Orange indicates, in this dataset, the position of each sub-prefecture of Ivory Coast. It is helpful for studies using the individual long-term traffic dataset, described in Sec. 2.2.3, including information about calling activities of customers at a sub-prefecture level. We show in Tab. 2.3 an extract from this dataset.

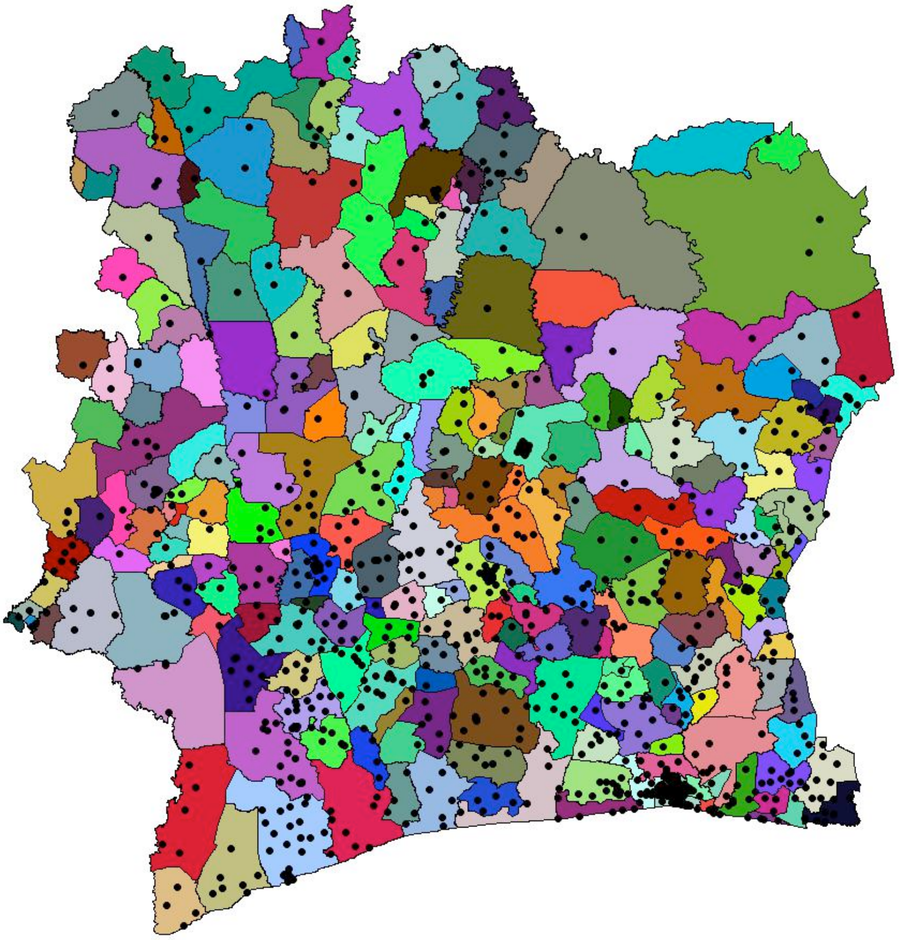


FIGURE 2.2: Positions of Orange's antennas in Ivory Coast and sub-prefectures administrative regions, taken from [18].

2.2.3 Traffic datasets

In the following, we describe the different mobile traffic datasets extracted by Orange from their CDR database. We note that, due to technical data collection problems, some information is missing from the datasets. In fact, the antenna identifiers are absent for a quarter of the recorded mobile phone activities. Corresponding sub-prefecture identifiers are also absent. In such cases, Orange replaces the antenna and sub-prefecture identifiers with -1.

Aggregate antenna-to-antenna traffic dataset. For each couple of antennas, this dataset provides the total number and the total duration of hourly exchanged calls among users located within the corresponding cells. This dataset spans over the whole observation period and excludes communications originating from, or destined to customers of other operators. Files providing this information are presented according to the format in Tab. 2.4.

This dataset can be employed for studies aiming at characterizing dynamic features of human behaviors. For urban and rural planning studies, it can be used to investigate correlations between

date_hour	originating_ant	terminating_ant	nb_voice_calls	duration_voice_calls
2012-04-28 23:00:00	1236	786	2	96
2012-04-28 23:00:00	1236	804	1	539
2012-04-28 23:00:00	1236	867	3	1778
2012-04-28 23:00:00	1236	939	1	1
2012-04-28 23:00:00	1236	1020	6	108

Table 2.4: Extract from aggregate antenna-to-antenna traffic dataset.

user_id	date_time	antenna_id
437690	2011-12-10 10:51:00	980
316462	2011-12-10 16:12:00	607
277814	2011-12-10 20:48:00	560
419518	2011-12-10 10:05:00	-1
18945	2011-12-10 11:32:00	401

Table 2.5: Extract from individual high spatial resolution traffic dataset.

human activities and geographical space features. It can thus allow to explore direct relationships between land use and human activity patterns [11]. For networking studies, this dataset constitutes a powerful tool, which allows to understand aggregate usage patterns characteristics over the whole operator's access network, so as to extract spatio-temporal network-wide usage features. It can also be very useful for the evaluation of networking-related applications [12].

Individual high spatial resolution traffic dataset. Individual anonymized CDRs are provided in this dataset for 50,000 randomly sampled users from Orange's CDR database, over ten two-week periods. This dataset allows to follow one user for 14 days, with per-call location information approximated at the cell-level. An extract from the dataset is presented in Tab. 2.5.

Information obtained through this dataset can be valuable for various mobility and networking studies. Clearly, it can capture individual mobility features relating to spatio-temporal travelling characteristics, with particular benefits to modeling and predicting human mobility patterns [9, 21]. From a networking point of view, such information can be explored for the characterization of individual calling patterns, towards the improvement of the cellular network performance [22].

Individual long-term traffic dataset. This dataset captures the calling activities of 50,000 randomly chosen individuals, from Orange's CDR database, over the whole observation period with a coarse geographical localization information. Each time a user makes a call, his position is provided at the level of sub-prefecture to which belongs the antenna. This dataset is presented according to the structure indicate in Tab. 2.6.

user_id	date_time	subpref_id
134931	2011-12-02 10:50:00	60
89571	2011-12-02 10:49:00	39
457232	2011-12-02 16:05:00	60
155864	2011-12-02 09:26:00	60
280671	2011-12-02 13:24:00	-1

Table 2.6: Extract from individual long-term traffic dataset.

source_user_id	destination_user_id
1052	20002
20002	20022
20018	20019
1052	20019
20019	20030

Table 2.7: Extract from the individual communication subgraphs dataset.

Similarly to the previous dataset, the current one can also be applied for mobility- and networking-oriented studies. However, it can be particularly of higher utility for mobility studies requiring long-term tracking of individual activities, and tolerating low spatial granularities: as an example, this dataset can be useful for the analysis of nation-wide travelling behavior of individuals out of their region of residence.

Individual communication subgraphs dataset. In this dataset, communication subgraphs are provided for 5,000 randomly chosen users. They include, for each selected user, his second order neighborhood communications over periods of two weeks. This means that a subgraph reflects communications between a user and his contacts, as well as communications between each one of his contacts and their own contacts. In the subgraph, a node represents an individual. For a couple of nodes, an edge exists between them if a communication took place between the two, over the two-week period. A subgraph does not provide any information concerning the number of communications, total communication time or the direction of the communication. The dataset is given according to the format in Tab. 2.7.

The importance of this dataset relates to social studies aiming at investigating the evolution of individual communication graphs over time. However, its utility is strongly affected by the fact that information is aggregated over each couple of week, making it hardly exploitable.

2.3 Processing datasets towards networking studies

In the following parts of the manuscript, we employ the aggregate antenna-to-antenna traffic dataset. For our work, we focus on calling activity in urban environments. Moreover, for the

date_hour	antenna_id	nb_voice_calls	duration_voice_calls

TABLE 2.8: Extract from the generated files with aggregate calling information.

evaluation of our networking applications, we need to acquire information concerning the demand and mobility of users over fine-grained temporal scales. Thus, in a first step, we perform pre-processing operations of data aggregation and filtering over the initial dataset. In a second step, we perform additional processing that depends on the scope of the study. We detail these steps in the following.

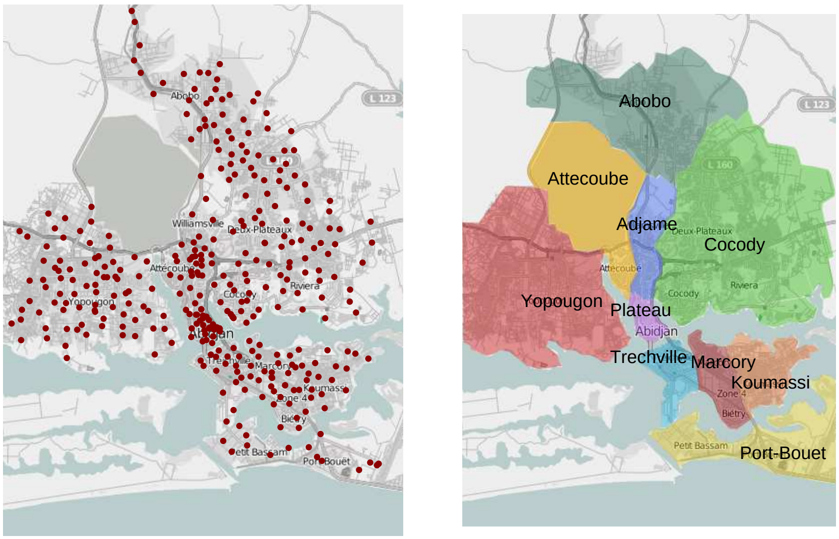
2.3.1 Pre-processing steps

Per-cell traffic aggregation. As explained in the previous section, the aggregate antenna-to-antenna traffic dataset provides, for each couple of antennas in Ivory Coast, the hourly number and duration of exchanged calls between them. As a first pre-processing step, we aggregate the hourly incoming and outgoing calling activity information over each antenna. For our work, this allows us to capture the global activity over the system. We thus generate new files with the structure in Tab. 2.8.

Urban data extraction. Our work targets urban environments, more complex to manage than rural ones from a networking perspective. In particular, we consider the case of Abidjan, the economical capital of Ivory Coast, a highly populated city with more than four million inhabitants over an area of 422 km^2. Hence, we filter the country-wide data by preserving only information relating to antennas in Abidjan. This leaves us with information concerning 364 antennas, out of the total 1231 Orange antennas deployed in the whole country. Thanks to this dense infrastructure deployment, the Abidjan data allows to capture human dynamics with a high level of precision, while keeping a good level of flexibility for the exploration of this data. Fig. 2.3(a) and Fig. 2.3(b) portray the position of antennas over the map of Abidjan and the geographical span of the communes of the city, useful for our work, as we explain later.

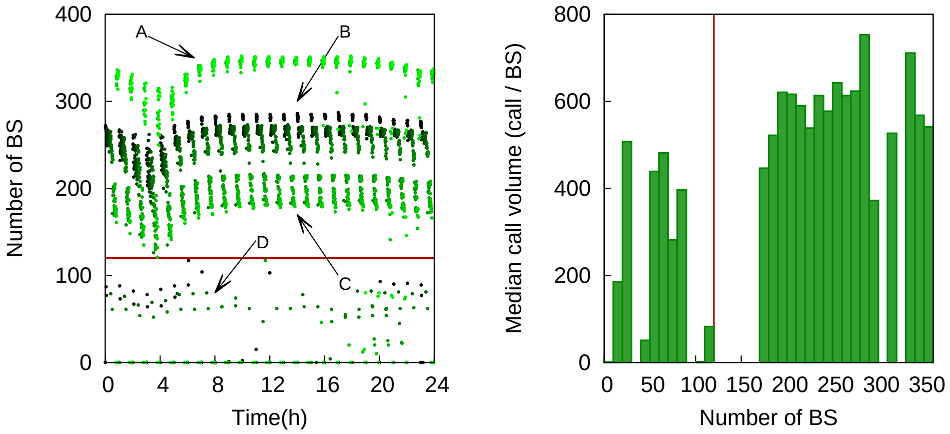
2.3.2 Characterization-related processing

As mentioned in the previous section, some data is missing from the dataset over a set of antennas, due to problems encountered by Orange, as well as electricity failures occasionally occurring in Ivory Coast. To avoid possible biases on our dataset analysis, we drop information corresponding to such situations, by examining the reliability of hourly collected information. In Fig. 2.4(a), we plot the evolution of the number of operational antennas for each hour over the five months: the x coordinate maps to the hour of the day, while the y coordinate provides the number of antennas over which non-zero traffic is recorded. We distinguish between the



(a) (b)

Figure 2.3: (a): Geographical distribution of antennas in Abidjan. (b): Communes of Abidjan.



(a) Number of BSs (b) Median call traffic

Figure 2.4: (a): Number of active antennas per hour for each day of the five-month dataset. The color contrast degradation allows to distinguish between different days, such that dark green maps to the first day in the dataset and light green maps to the last one. (b): Median volume per base station as a function of the number of active BSs, aggregated over the 5-month period for hours between 10:00 and 20:00.

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.

Three major situations are detected. The first one, labeled as A in the figure, includes the highest number of BSs, around 350, and goes from March 28th, 2012 until April 22nd, 2012. The second situation, tagged as B, with almost 250 antennas, spans between December 7th, 2011 and February 21st, 2012. Situation C, with 170 BSs, goes from February 22nd, 2012 until March 27th, 2012. Fluctuations also appear in each of these behaviors, with local minima emerging in the night hours between times 3:00 and 6:00, while the number of antennas stays almost stable for the rest of the day. These minima are due to the reduction of cellular traffic at night hours with a very small number of individuals active over the network.

Additionally, a particular behavior, referred to as D, with a very small number of antennas, is detected for short intervals dispersed over the entire five-month period. These are the result of missing information over a majority of BSs in the city and can be related to one of the following two situations. The first possibility is that information is missing for most of the antennas, but the remaining ones still present their typical traffic, reflecting a data collection problem only for absent antennas. The second possibility is that there is a global problem in the network, affecting the collection process and/or the actual usage behaviors over the city-wide access network, leading to irregular recorded data over present antennas.

We thus investigate the reliability of such situations, in terms of call traffic information, in Fig. 2.4(b), by plotting the median call traffic volume per BS with respect to the number of BSs appearing over the one-hour time interval. More precisely, each bar in the figure represents the median call volume per BS, obtained when considering one-hour time intervals with the same number of BSs specified on the x axis. Information portrayed there is limited to day hours with significant traffic, i.e. between 10:00 and 20:00, over all the observation period, so as to avoid any bias implied by low-traffic night hours.

We note that the intervals centered around the values of 305 and 325 antennas present null median volume values due to the fact that no information is recorded over one hour with a number of BSs lying in one of these intervals. We can observe from Fig. 2.4(b) that behaviors A, B and C are characterized by high median volume per BS, while for behavior D, the median call volume oscillates over a significant interval. This shows that the second possibility holds for our case, confirming the fact that information is missing from the dataset due to a global problem in the network. Thus, we exclude from our analysis all the intervals falling in the D behavior, i.e. providing information for less than 120 antennas.

2.3.3 Exploitation-related processing

Hourly aggregated information cannot be directly employed for our networking-oriented exploitation of datasets. In fact, both our applications include network resources management procedures, which require very precise spatio-temporal consumption information, lacking from the D4D datasets. Moreover, our mobility-driven application requires a detailed view of users movements and, in particular, handovers they encounter as they traverse network cells. In the following, we describe how we generate the corresponding information.

2.3.3.1 Per-frame traffic generation

Depending on the objective of the study, the level of granularity required for the exploitation of datasets can vary. For some, it can be more beneficial to explore information aggregated over relatively long periods. As an example, for studies aiming at detecting activity hotspots, so as to increase network capacity through localized network densification, considering only snapshots of traffic consumption over small durations may lead to erroneous detections of hotspots. Instead, information aggregated over long time intervals is more reliable from this point of view, reflecting more stable behaviors. On the other hand, some studies may require fine-grained data. This is typically the case for works implying management of network radio resources. Both of our target networking applications demand such mechanisms, which leads us to performing an additional processing step in order to generate consumption information over significant time intervals.

As the D4D data reflects users communications over Orange's 2G network, we derive information over each GSM frame, the smallest time unit for handling network resources in GSM systems. To do so, we carry out three successive steps for data spanning over one working day: *i*) Starting frame attribution. *ii*) Call duration assignment. *iii*) User location selection.

Starting frame attribution. In this first step, we attribute to the hourly calls at a BS their starting frames. We consider that the number of calls arriving during one frame follows a Poisson distribution with parameter λ equal to the average number of calls that arrive per frame. More precisely, λ is equal to the total number of calls over an hour at a specific BS of interest divided by the number of frames per hour. We note that a GSM frame lasts for approximately 4.615 ms.

Call duration assignment. As a second step, we assign to each call its duration according to a Log-normal distribution [23]. The location μ and scale σ parameters of the distribution are derived based on the mean m and standard deviation s of the non-logarithmized hourly calls duration as follows:

$$\mu = \ln(\frac{m^2}{\sqrt{s^2 + m^2}})$$

and

$$\sigma = \sqrt{\ln(1 + \frac{s^2}{m^2})}$$

In our case, we are only able to obtain m from the dataset, as we only have the information concerning the aggregate duration of calls and their number over each hour. We observe that its value can go up to 154 s. As for the standard deviation s, we suppose its value is equal to 1 s.

	Plateau	Treichville	Marcory	Koumassi	PortBouet	Adjame	Cocody	Yopougon	Attecoube	Abobo
h	40	12	16	12	4	16	16	12	8	8
d	10	5	3	3	3	5	10	3	3	3

TABLE 2.9: Average building height (h) and average separation distance between buildings (d) for each commune. These values are obtained based on evaluations over google maps.

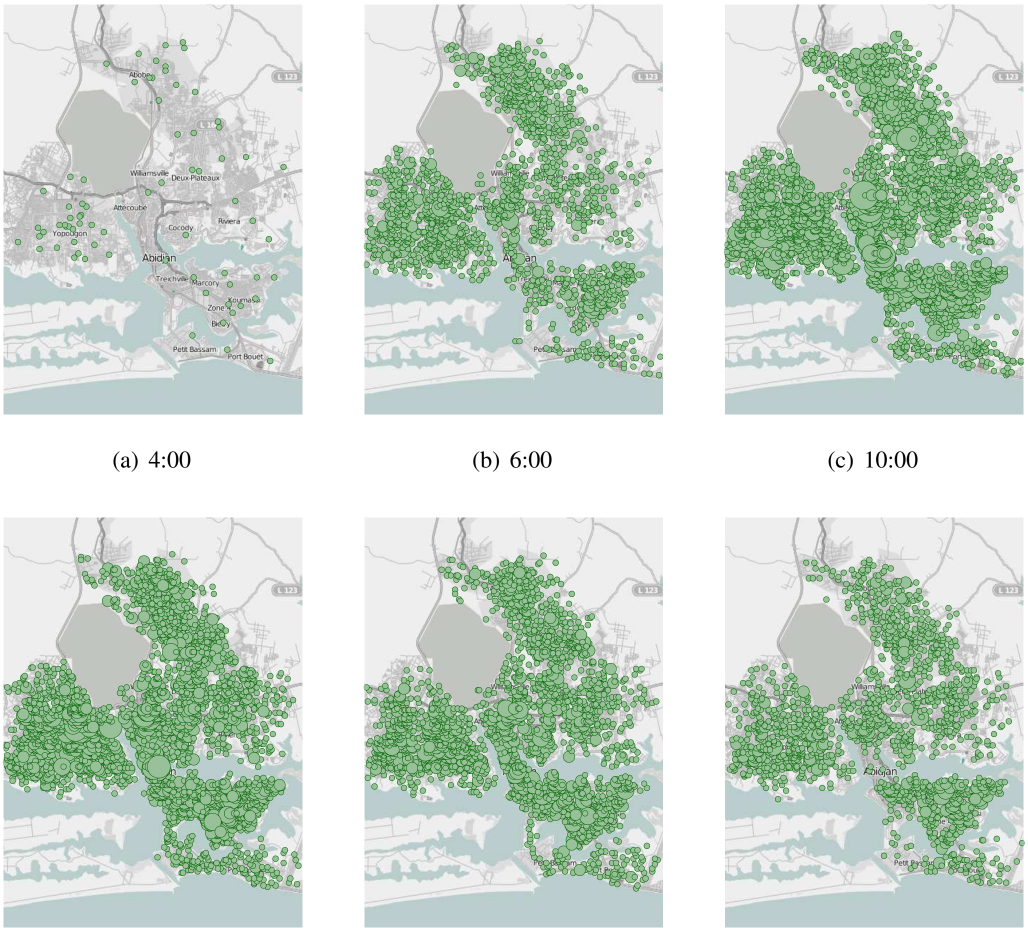
User location selection. In our third step, we attribute to each call its location over the discretized space. We consider that space is divided into small squares of 100 m by 100 m each. We assume that calls recorded at a BS are uniformly distributed over the set of small areas it covers when transmitting at maximum power. Thus, for each call arriving at a BS, we randomly pick an area that the BS can cover, and consider that the call originates from there.

Eventually, the set of areas covered by a BS depends on the signal propagation model employed. In our case, we use the Walfish-Ikegami empirical model [24], which provides the average path loss between a transmitter and a receiver, taking into account the properties of the vertical plane between them, through a set of input parameters. We assume that an area is covered by a BS if the user appearing at the farthest edge of the area, with respect to the BS, receives a minimum power of -100 dBm. Moreover, we consider that a BS can transmit at a maximum power of 20 W and operates over the 900 MHz band.

We remark that the communes of Abidjan are diverse from a regional planning point of view, which can affect the shape of signal propagation there. As an example, very tall buildings dominate over the city center, while residential areas are mostly formed by buildings of one or two floors. We thus tune the input parameters of the propagation model separately for each commune, according to the average characteristics of the environment there. In particular, the average building height h and the average separation distance between buildings d are chosen according to Tab. 2.9. We consider that the width of road w and the road orientation with respect to the direct radio path h are constant over the various communes, with values equal to 10 m and 70°, respectively. Additionally, we choose the BS height as equal to 30 m over all communes, except for the Plateau, where we consider it as equal to 50 m. Finally, we assume that mobile devices are present at a height of 1.7 m.

By the end of this step, we are able to propose information concerning ongoing calls for each GSM frame over the day. In our evaluations, we focus on six hourly frames, separated by time intervals of 10 minutes. We show in Fig. 2.5 a geographical representation of the generated ongoing calls for each area for an extract of frames over the day. Each circle covering an area represents the number of ongoing calls there. We can notice these plots reflect the typical daily traffic behavior, with denser plots including larger circles appearing at high activity hours.

To verify that our additional processing steps preserve the major features of the dataset, we derive the main characteristics of both the initial and generated datasets in terms of total number and duration of calls per hour for each BS. Fig. 2.6 and Fig. 2.7 portray the Probability



(a) 4:00

(b) 6:00

(c) 10:00

(d) 14:00

(e) 18:00

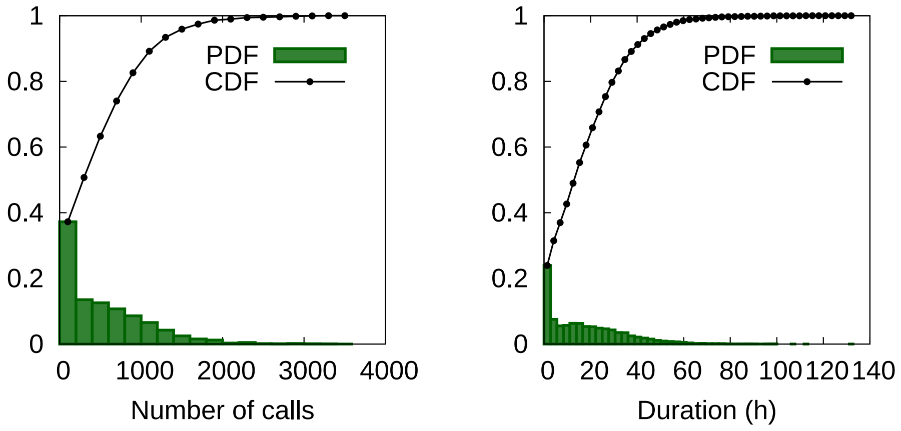
(f) 22:00

FIGURE 2.5: Number of calls per BS for an extract of frames over the day.

Distribution Function (PDF) and Cumulative Distribution Function (CDF) for the two datasets respectively. We can notice the distributions are quite similar.

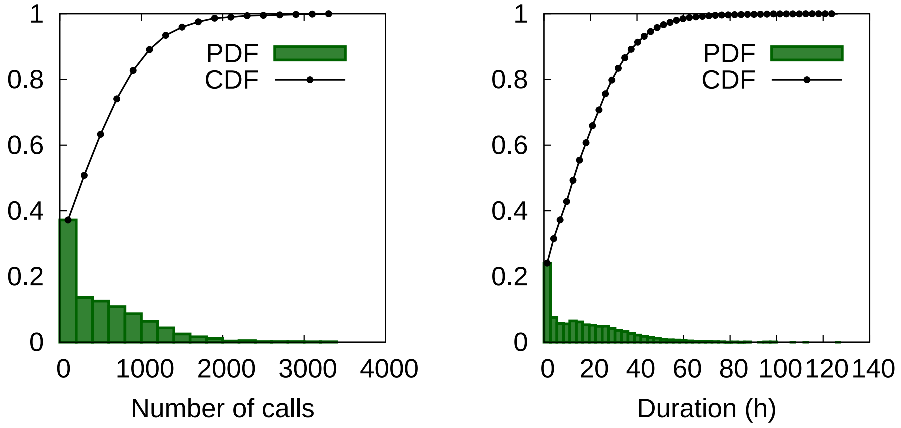
We also perform a χ^2 test [25] to compare the two distributions. The χ^2 test consists of testing the null hypothesis of independence between two variables. It relies on the calculation of the χ^2 statistic value [1] and the corresponding p-value representing the probability of observing a sample statistic as extreme as the obtained χ^2 statistic value. The null hypothesis is rejected if the p-value is less than a pre-defined significance level, typically equal to 0.01 or 0.05. In our case, the test allows us to determine whether there is a significant association between the variables corresponding to the initial and generated datasets in terms of total number and duration of calls per hour for each BS. We obtain p-values of $2.e^{-}04$ and $1.7e^{-}05$ for the number and duration of calls respectively, which represents a strong evidence against the null hypothesis, leading to its rejection, and thus confirming the nice visual match between the distributions.

[1] Given two binned variables, with R_i the number of samples in bin i for the first variable and S_i the number of samples in the same bin i for the second variable, then the χ^2 statistic value is: $\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i}$.



(a) (b)

FIGURE 2.6: Initial dataset characteristics: Distribution of the total (a) number and (b) duration of calls per hour for each BS.



(a) (b)

FIGURE 2.7: Generated dataset characteristics: Distribution of the total (a) number and (b) duration of calls per hour for each BS.

2.3.3.2 Users mobility representation

The aggregate antenna-to-antenna traffic dataset yields records relating only to the initiation and termination of calls. They do not provide any information concerning signaling messages exchanged over the network at a regular basis, e.g. for broadcasting network-related information, or those complementing specific events occurring in the network, e.g. paging procedures. Acquiring such a knowledge can be very useful for studies aiming at evaluating and improving the performance of cellular network architectures, management functions and communication protocols.

In the particular case of our mobility-related application, a network-wide picture of handovers occurring between each couple of neighboring cells is needed with a precise temporal granularity. To generate this information, we proceed as follows. First, we consider that x% of the total calls experience a handover. We then randomly pick these calls among the total set of calls. Finally, we choose the destination cell according to a uniform distribution over the set of originating cell's neighbors.

Such spatio-temporal uniform distribution assumptions may be too strong from a mobility point of view. In terms of temporal properties, people are generally more mobile at some hours of the day comparing to others, e.g. a high level of mobility characterizes the morning hours as people go to work, with respect to evening hours during which people tend to stay at home. Similar observations also hold from a spatial perspective, i.e. people's macroscopic mobility flows tend to follow regular major directions. As an example, typically significant macroscopic mobility flows are directed towards working areas in the morning, as opposed to those representing the reverse afternoon behavior as people move back to residential areas.

However, understanding people's mobility patterns is not sufficient to acquire a solid knowledge on handovers occurring in the network, as it needs to be combined with users calling behavior, i.e. what we precisely need is a clear illustration of users movements when they are communicating over the network. Eventually, acquiring such a detailed information is not an easy task and requires analyzing traces recorded at the level of RNCs, which we do not possess. Nevertheless, we employ in our study two handover scenarios, with the assumptions of 5% and 50% of total calls encountering a handover, which can be representative of two extreme cases: low and high mobility scenarios.

2.4 Summary

In this chapter, we present general aspects relating to mobile traffic datasets. We provide a global overview of their collection process. In particular, we describe cellular networks architectures, and present probes used to gather the data. We also outline main characteristics of these datasets. We then describe the datasets that we employ in this thesis and explain how we process them towards networking studies.

Part II

A Networking Perspective on Mobile Traffic Datasets Analysis

Chapter 3

Analyses of mobile traffic datasets

3.1 Introduction

Mobile traffic datasets hold a very promising potential when it comes to cellular network systems analysis. They can capture the dynamics of mobile traffic over time and space, at unprecedented very large scales. From a networking perspective, this is of particular importance to understanding and characterizing the evolution of traffic consumption on the cellular access network, paving the way towards more efficient and performant user-oriented networking solutions.

In this chapter, we review previous works analyzing the spatio-temporal evolution of mobile traffic. Our objective is to provide the reader with an overview of the main tools employed in these studies and provide a synthesized outline of major results they derive. Thus, we start by introducing important concepts and techniques applied in analyses of mobile traffic datasets, followed by a discussion of the findings of network-oriented analyses, from two different perspectives: *i*) An aggregate point of view, where the overall traffic, relative to all users accessing the network within a certain geographical area, is analyzed; *ii*) An individual mobile user point of view, where the behavior of each customer or mobile device is accounted for by itself.

We summarize the works discussed in this chapter in Tab. 3.1, and present the main features of the studied datasets therein. Overall, we note that the datasets are quite diverse. While operators are not specified for a significant part of works, for privacy concerns, we still notice a variety of providers, when known. Even more, we can observe the datasets emerge from different regions, spread all over the world, with information about thousands and even millions of users, covering long durations, that can go up to a whole year. The heterogeneity of datasets is also observed in terms of traffic type. There, we notice voice, texting and data activity are recorded, although the first two dominate over the majority of works. In addition to the features of datasets, Tab. 3.1 highlights the network-related research aspects for each paper, which we discuss in more detail in this chapter.

	Analysis		Dataset							Focus						
	Name	Date	Operator	Area	Time	Users	V	T	D	TD	SD	SP	AD	UC	TM	DT
Aggregate	Williamson [26]	11/05	–	100 cells	1 week (2004)	10 K			✓	✓			✓		✓	
	Keralapura [27]	09/10	–	USA	1 day (2008)	500 K			✓	✓				✓		
	Paul [28]	04/11	–	One country	1 week (2007)	100 K			✓	✓	✓		✓		✓	
	Shafiq [29]	06/11	–	One state	1 week (2010)	~ M			✓	✓			✓			✓
	Zhang [30]	08/12	–	–	1 week	50 K			✓	✓						✓
	Mucelli [31]	09/14	–	Mexico city, Mexico	1 week (2013)	2.8 M			✓	✓			✓	✓		
	Wang [32]	04/13	–	2 cities	Months (2007/11)	2.4 M	✓	✓		✓						
	Girardin [33]	06/09	AT&T	NY, USA	1 year (2007/08)	–	✓	✓		✓	✓	✓				
	Hohwald [34]	06/10	–	Metropolis	6 months	50 K	✓	✓		✓						
	Cardona [35]	12/14	–	European country	7 months (2011/12)	40 K			✓	✓						
	Pulselli [36]	06/08	Telecom Italia	Milan, Italy	2 months (2004)	–	✓				✓					
	Girardin [37]	10/08	Telecom Italia	Rome, Italy	3 months (2006)	–	✓	✓			✓					
	Shafiq[4]	12/13	–	USA	1 week (2010)	–			✓		✓					✓
	Ratti [38]	11/06	–	Milan	2 weeks (2004)	–	✓	✓			✓					
	Willkomm [39]	10/08	–	NC, USA	3 weeks	–	✓				✓		✓			
	Csáji [40]	06/13	Orange	Portugal	–	100 K	✓				✓					
	Cerinsek [41]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓			✓			✓		
	Hoteit [19]	12/12	Orange	Paris	2 days (2012)	>1.5M			✓		✓					
	Shafiq [20]	03/12	–	Metropolis	32 hours (2010)	~ 10 K			✓		✓					
	Almeida [42]	09/99	Telecel	Lisbon	3 days (1997)	3 M	✓	✓			✓					
	Soto [11]	06/11	Telefonica	Madrid and Barcelona	1 month (2009)	3 M	✓	✓			✓					
	Cici [43]	06/15	Telecom Italia	Milan, Italy	1 month (2013)	–	✓	✓			✓					
	Vieira [44]	08/10	Telefonica	2 metropolis	4 months	1 M	✓				✓					
	Trestian [45]	11/09	–	5000 km^2	1 week	281 K			✓		✓				✓	
	Trasarti [46]	05/13	–	Paris, France	–	–	✓	✓			✓					
	Zong [59]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓			✓					
	Xavier [47]	12/12	Oi Telecom	Rio de Janero, Brazil	3 days (2011)	–	✓	✓				✓				
	Shafiq [48]	06/13	-	2 metropolis	Several days (2012)	100 K	✓		✓			✓				
	Paraskevopoulos [49]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓				✓				
	Gowan [50]	05/13	Orange	Ivory Coast	5 months	5 M	✓	✓				✓				
	Xavier [51]	05/13	–	4 cities (Brazil)	4 days (2011/12)	–	✓	✓				✓				
	Pastor-Escuredo [52]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓				✓				
	Elzen [53]	05/13	Orange	Ivory Coast	5 months (2012)	–	✓	✓				✓				
	Dixon [55]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓				✓				
	Altshuler [58]	08/13	–	European country	3 years	–	✓	✓				✓				
Individual	Candia [54]	07/08	–	230400 km^2	–	–	✓					✓	✓		✓	
	Dasgupta [72]	03/08	–	–	5 months (2007)	3.1 M	✓						✓			
	Ben Abdesslem [73]	03/14	–	European country	8 weeks (2011/12)	3 M			✓				✓			
	Lin [70]	10/07	–	Northern PRC	–	600 K	✓	✓						✓		
	Becker [71]	06/11	–	Morristown, USA	2 months (2009/10)	475 K	✓	✓						✓		
	Couronné [74]	10/11	Orange	Paris, France	1 day	4 M	✓	✓							✓	

TABLE 3.1: Main features of works analyzing mobile traffic data towards understanding resource consumptions. In the analysis columns, date is in MM/YY format. In the dataset columns, V is voice, T is texting, D is data. In the focus columns, TD is traffic temporal dynamics, ST is traffic spatiotemporal dynamics, SP is special dynamics, AD is activity distributions, UC is users categories, TM is traffic-mobility correlations, DT is device and traffic types.

3.2 Tools and methods

This section introduces a set of tools and methods that are widely employed for the analysis of mobile traffic studies. We cover the main approaches used to represent mobile traffic datasets, as well as common techniques and algorithms used to process them.

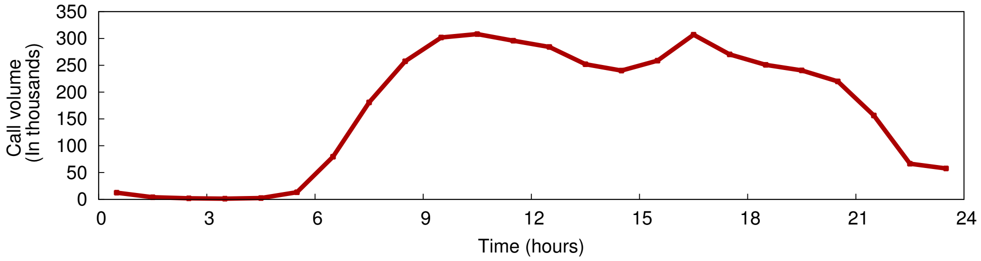


FIGURE 3.1: Mobile traffic time series representation of the aggregate hourly voice traffic recorded by Orange, over its network in the city of Abidjan, on April 3rd, 2012.

3.2.1 Representation of mobile traffic datasets

Two major approaches have been adopted in the literature for the representation of mobile traffic datasets, as detailed next.

Mobile traffic time series: The evolution of mobile traffic is very often modeled with a time series representing a set of traffic measurements over regular time intervals [4, 11, 19, 20, 26–42, 44–55]. For networking analysis, this representation is especially useful for tracking the temporal evolution of both aggregate and individual traffic. Additionally, it is applied with various levels of granularity, and diverse kinds of measurements, that can provide information concerning total or directional activities, of voice, texting and data traffic. Fig. 3.1 shows an example of a mobile traffic time series representing the aggregate hourly voice traffic recorded by Orange, in Abidjan, on a typical working day in 2012.

Mobile call graph: While time series allow to capture the temporal evolution of traffic, they do not allow to account for interactions between various entities in the network. To tackle that, several studies employ a graph-based model $G(\mathbb{V},\mathbb{E})$ [5] which describes interactions between a set of nodes, or vertices $\mathbb{V}$, through a set of edges $\mathbb{E}$ linking them. This representation allows to exploit analyses techniques from the field of graph theory, so as to infer fundamental properties of the mobile call graph [56, 57].

The graph structure has been applied for individual-based analysis [58], where $\mathbb{V}$ represents a set of users and $\mathbb{E}$ a set of exchanged traffic activities among them. This definition is especially of interest for studies aiming at characterizing the way users interact. The graph structure can also be used to study interactions among infrastructure entities. It has been constructed over base stations forming the set of vertices $\mathbb{V}$, with edges modeling the cellular traffic exchanged between couples of base stations [44, 59]. We show, in Fig. 3.2, an example of such a mobile call graph, representing major interactions among Orange's base stations in Abidjan, on December 10th, 2011. The size and the color of a node refer to the overall level of activity over each base station: the darker and the bigger the node is, the higher is the traffic recorded there.

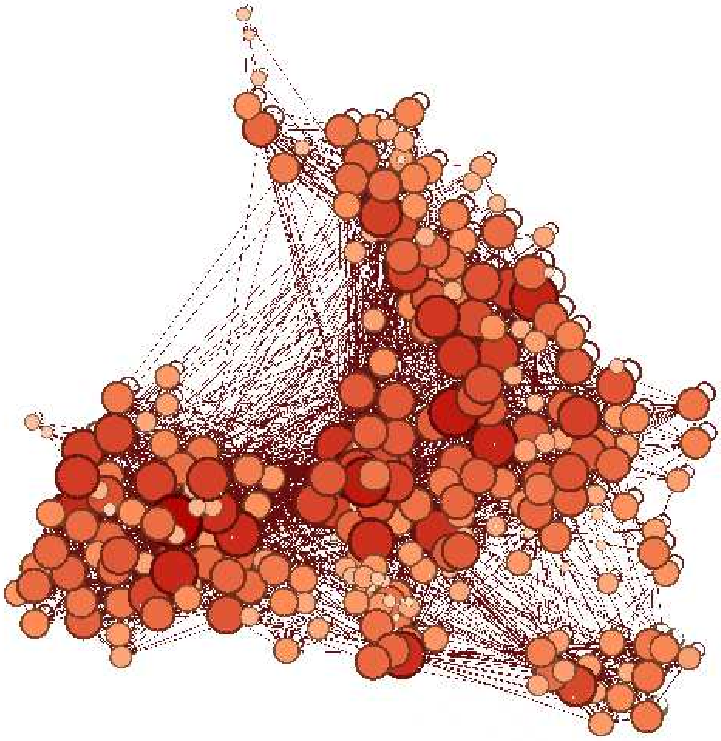


FIGURE 3.2: Mobile call graph representation of major interactions among Orange base stations in the city of Abidjan, on December 10th, 2011. Nodes represent base stations, whose color and size refer to the level of activity recorded there, such that darker and bigger nodes reflect a higher traffic. Edges represent important exchanges of traffic between couples of base stations.

More detailed graph structures have also been used [56, 57, 60–62]. In some studies [56, 60, 61], edges are assigned weights, allowing to quantify the level of interaction between pairs of nodes. Additionally, directed edges have been employed [57, 62], to distinguish between in-coming and out-going activities. Temporal graph representations, remain less common [56], although carrying very promising potential towards understanding the dynamics of interactions in the network.

3.2.2 Clustering methodologies

Clustering techniques constitute fundamental tools in the field of data mining [63]. Due to their well-known high capabilities to infer relationships between data objects, they have been largely employed by researchers for the analysis of mobile traffic datasets.

Clustering strategies aim at dividing objects into clusters or categories, such that similar items are grouped together, and dissimilar items are placed in different categories. Clustering algorithms operate over a set of vectors $\mathbb{X}$, where a vector $\mathbf{x}_i \in \mathbb{X}$ represents an object i. We refer to a feature j of $\mathbf{x}_i$ as x_i^j. Clustering techniques rely on distance measures, or alternatively similarity measures, in order to build categories of objects. As an example, the euclidean distance, computed between the object vectors, is one of the main metrics used to quantify the level of dissimilarity between objects. In the following, we use d_{ij} to denote the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Clustering strategies can either adopt a hierarchical or a partitional approach [63]. Hierarchical clustering techniques organize objects into a hierarchical structure, that can be visually represented with a dendrogram. Hierarchical algorithms can follow an agglomerative method or a divisive one. Agglomerative algorithms start from singleton clusters, each including only one single object, and at each iteration, merge clusters according to a certain function, leading finally to one cluster with all objects belonging to it. Classical examples include: single linkage [64], average linkage [65], and Ward's method [66], which adopt the same approach, but differ in their cluster merging decision function. Divisive algorithms, e.g. DIANA [67] and MONA [67] algorithms, follow the reverse approach, starting from one big cluster, and dividing it in singletons. However, they are generally more computationally expensive than agglomerative algorithms. On the other hand, partitional clustering algorithms, such as k-means [68] and ISODATA [69], directly group objects into a number of categories k. An important remark is that hierarchical algorithms can also be used to group objects into a fixed number of categories. This can be completed by stopping the algorithm at the required level. In all cases, there is no agreed rule to define the exact number of categories, its choice remains either determined based on some knowledge about the data, or according to some clustering quality measures.

Next, we provide more details about the main algorithms applied in the analysis of mobile phone datasets: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [31, 43, 49] and k-means [11, 20, 30, 31, 39, 41, 70, 71].

UPGMA clustering method: The UPGMA hierarchical clustering algorithm, also known as mean or average linkage clustering algorithm, is an agglomerative hierarchical clustering technique, which starts from having each object in a separate cluster, and merges at each iteration the two clusters that show the highest level of similarity on average, as follows. Suppose that, at iteration n of the algorithm, objects are placed into disjoint clusters $\mathbb{C}_k^n$ forming a set $\mathbf{C}^n$. The algorithm computes the average distance between each pair of clusters $\mathbb{C}_k^n$ and $\mathbb{C}_h^n$ in $\mathbf{C}^n$ as:

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

It then merges the two clusters $\mathbb{C}_k^n$ and $\mathbb{C}_h^n$ implying the smallest value of d_{kh}^n into a new cluster $\mathbb{C}_m^{n+1}$, for the following $n+1$ iteration. Finally, the set $\mathbf{C}^{n+1}$ is defined accordingly.

***k*-means clustering algorithm:** The k-means algorithm is the most used partitional clustering algorithm. It aims at grouping objects in $\mathbb{X}$ into a set of k clusters $\mathbb{C}$, with the objective of minimizing the within-cluster distances. The algorithm operates as follows. It starts by initializing a k-partitioning of objects, based on random selections or some prior knowledge. Then, it evaluates the centroid $\mathbf{m}_k$ of each cluster $\mathbb{C}_k \in \mathbb{C}$ as follows:

$$\mathbf{m}_k = \frac{1}{|\mathbb{C}_k|} \sum_{i \in \mathbb{C}_k} \mathbf{x}_i$$

In an iterative process, the algorithm places each object in the closest cluster to it and recalculates the centroids according to the new configuration. The algorithm repeats the last step as long as there remains possible changes for any cluster.

3.3 Aggregate traffic characterization

Analyzing the spatio-temporal evolution of global consumption of users has been the focus of a large number of previous studies. These analyses aim at understanding the evolution of the aggregate demand over the access network, in its typical variations and outlying behaviors.

3.3.1 Temporal dynamics

Many works aimed at investigating the temporal evolution of mobile traffic by employing the mobile time series representation. Accordingly, mobile traffic is observed to follow a regular behavior. While analyzing the packet data call activity over 100 cells for a duration of one week, Williamson *et al.* [26] notice that the traffic presents a daily repetitive behavior over weekdays, characterized by a low activity during the night and a high activity during the day. The same behavior is also perceived at various scales in several studies by Keralapura *et al.* [27], Paul *et al.* [28], Shafiq *et al.* [29], Zhang *et al.* [30], and Mucelli *et al.* [31]. In particular, Paul *et al.* [28] capture this regularity at a nation-wide scale, as they derive the distribution functions of daily traffic and its temporal evolution over a week.

Although the most significant load difference is between night and day, some variability in the mobile traffic can still be noted over smaller durations. Wang *et al.* [32] detect variations at an hourly basis, of both calling and texting activities in two different scenarios: San Francisco, and an unnamed Chinese city. Interestingly, they identify the occurrence of two daily peaks over the Chinese dataset for calls and one peak for texting, while in the case of San Francisco, they detect only one peak for calls and no peak at all for texting. Williamson *et al.* [26] detect temporal variations at an even finer granularity, of 10 minutes, in packet data call traffic. They detect the presence of several peaks, with the largest ones appearing mostly in the late afternoon.

Significant differences are also observed for various traffic types between weekends and weekdays. Williamson *et al.* [26] notice that packet data call traffic is higher on weekdays with respect to weekends. Other works by Zhang *et al.* [30], Wang *et al.* [32], Girardin *et al.* [33] and Hohwald *et al.* [34] confirm these observations. However, the latter authors underline the fact that Sunday calls last longer than others, as they study the calling behavior of mobile users in a metropolitan area for a duration of six months.

Finally, Cardona *et al.* [35] capture seasonal variations in users traffic: they detect an increase of 20% in monthly data usages towards the end of the year with respect to data usages in the summer.

From a technical perspective, Shafiq *et al.* [29] focus on the possibility of predicting the temporal evolution of network load. They introduce a simple, yet powerful, Markovian model which allows to predict future aggregate traffic state, based on its past evolution. Their model separates between variations over weekends and weekdays, and is parameterized according to some measured traffic statistics. The evaluation of their strategy over a country-wide dataset, covering several days, shows that they are able to accurately predict future evolution of traffic, with a significantly low mean squared error of $1.7e^{-4}$.

Key points. The analyses of the temporal evolution of aggregate traffic have unveiled a repetitive daily binary-like behavior characterized by a low traffic at night hours and a high traffic during day hours, across different datasets. Nevertheless, some studies have pointed to traffic variations over different hourly, weekly and monthly temporal scales, suggesting the need for further investigations to understand the originating human activity dynamics.

3.3.2 spatio-temporal dynamics

While studying the temporal evolution of the aggregate traffic can uncover important characteristics of the consumption patterns, it does not allow to understand how usages vary over different areas, a very useful analysis for networking studies.

In a seminal work, Girardin *et al.* [33] consider the evolution of traffic over different areas of interest in New York. They observe that these regions exhibit a similar average behavior during the working days, while they present a clear variability during the weekends. The difference in the spatial distribution of mobile traffic between working days and weekends is also indicated by other works. By representing aggregate daily activity over Milan with geographical plots, Pulselli *et al.* [36] notice that, on weekends, the calling activity is concentrated in peripheral residential areas, while, during the weekdays, the calling activity is concentrated in the city center. Girardin *et al.* [37] also detect such discrepancies in Rome, characterized by a high level of activity during weekends around the Colosseum, reflecting the presence of tourists in the area, and a high calling activity close to the train station during weekdays, due to business visits to the city. On a finer temporal scale, they also detect variations in the traffic behavior during the evening hours, while the consumption remains quite similar among different regions over the rest of the day. Other works led by Paul *et al.* [28], and Shafiq *et al.* [4] later confirm these observations.

Several other works explore the more precise resolution granted by base stations and cells. Focusing on the daily traffic evolution in Milan over 16 days, Ratti *et al.* [38] show that some base stations are characterized by a high level of activity during the evening, while others present a high level of activity during office hours. More interestingly, they map the cellphone activity over the city and observe how the activity moves from the suburbs towards the city center between 9 AM and 1 PM.

Willkomm *et al.* [39] distinguish between three different representative cell behaviors, using *k*-means clustering algorithm over cells traffic patterns, in three urban areas in Northern California, for a duration of 3 weeks. They identify cells with low traffic at nights for all weekdays, cells with low traffic at nights for all weekdays and at daytime for weekends, and finally, cells with low traffic during all times. Three classes of base stations are also identified by Csáji *et al.* [40]. Using weekly time series, the authors map the detected categories to base stations in home, work and other locations. Finally, Cerinsek *et al.* [41] find five classes of base stations with similar daily and weekly traffic profiles, by combining two clustering algorithms: *k*-means method and Ward's hierarchical clustering algorithm. They also show that three of these classes present geographical correlation, as the corresponding base stations are located in close proximity.

Such observations are also detected based on application usages. Hoteit *et al.* [19] confirm the heterogeneity at the level of cells. By observing traffic demand over one day in Paris, they realize that some cells are more loaded than others, in terms of number of data users and applications traffic volume. Also focusing on application usages in a metropolitan area, Shafiq *et al.* [20] notice that the usages of the most popular applications, i.e. web browsing and email, are not uniformly distributed over space when considering data consumption over 32 hours. More generally, by applying the *k*-means algorithm over the behavior of cells, they conclude that the application usages depend on the location at both macro and micro levels.

Other studies consider an even deeper analysis, aiming at investigating underlying land use characteristics that can influence individual base stations consumption. In an early study, Almeida *et al.* [42] group base stations in Lisbon according to the land use of the area, with the purpose of modeling the mean temporal call variations. Their study shows that a trapezoidal function best fits the temporal evolution of traffic in residential and suburban areas, while a double gaussian model best fits base stations at major transport arteries.

Soto *et al.* [11] adopt a reverse approach to link the behaviors of base stations to different land use in Barcelona and Madrid. By employing *k*-means clustering algorithm over temporal series, they are able to distinguish major base station categories related to: work, residential, hybrid, nightlife and leisure activities. In a similar study, Cici *et al.* [43] apply an agglomerative hierarchical approach to cluster the behavior of base stations towards inferring the corresponding land use. They perform the clustering according to the decomposition of traffic time series into their frequency domain components. They show that their strategy allows to identify major clusters falling into the following land use categories: university, business, green and residential areas. More interestingly, they show that their strategy outperforms the one proposed by Soto *et al.* [11] according to the entropy measure, allowing them to assess the quality of clustering based on the density distribution of land use categories over the different clusters.

Vieira *et al.* [44] focus, instead, on the identification of hotspots in two metropolitan areas. By analyzing a graph describing interactions among neighboring base stations, they are able to detect dense activity zones at various times. Their results show that, on weekdays, base stations in the city center are highly loaded during the morning, while base stations in commercial and business centers accommodate heavy traffic loads during the rest of the day. In the weekend, they identify hotspots around commercial and business centers during the morning and the afternoon; and around commercial and night life areas during the evening and night hours. Similarly, Trestian *et al.* [45] identify day, noon, evening and night hostpots in a metropolitan region, and find them to be correlated to the nature of the geographical area they reside in.

Using a more complete approach, Trasarti *et al.* [46] investigate correlations between traffic variations over different geographical areas at successive instants. For this goal, they propose a method for the extraction of such patterns from mobile traffic data. One example that they detect is an increase in the activity at Charles de Gaulle airport followed by an increase in Gare de l'Est in Paris. Another particular methodology is also adopted by Zong *et al.* [59], who follow the temporal evolution of the base stations mobile call graph, with the objective of estimating the probability of having new important links in the graph, appearing at future time instants. The authors show that existing graph evolution models fail to capture the dynamics of the mobile call graph, and thus propose accordingly an adequate model.

Key points. There is a general agreement on the fact that mobile traffic presents heterogeneous patterns over space. This variability is observed to be related to the diversity in land use characteristics of the geographical space. However, these results stress the need for comparative studies to understand the impact of general societal lifestyles on the traffic consumption patterns.

3.3.3 Special dynamics

Human dynamics are affected by particular manifestations occurring in a certain region. These can be for example of social, political or economical nature. Several previous works compare network usages on special occasions to those on typical normal days, so as to infer the impact of special events on the traffic dynamics. Xavier *et al.* [47] focus on base stations covering the area of the stadium in Rio de Janero. They notice an increase in the number of calls before and after the match, accompanied by a decrease during a Sunday soccer match with respect to another Sunday with no match. Girardin *et al.* [33] also detect an increase in the calling activity of users over different areas of New York due to the Waterfall exhibition. They also point to the fact that some national festivities, such as Thanksgiving, Christmas, New Year's Eve, Easter and the 4th of July, imply a much higher absolute density of phone calls with respect to typical days. Similar observations are derived by Shafiq *et al.* [48], who illustrate how crowded sports and conference events lead to an increase in the access workload and result in significant voice and data performance degradation.

Paraskevopoulos *et al.* [49] propose a strategy to characterize the traffic patterns of base stations during special events by clustering their behavior using the UPGMA algorithm. By examining the case of Easter Monday, they notice that 39 cell towers witness an increase in the calling activity, while 1173 present a decrease.

Instead, Gowan *et al.* [50] study the duration of calls during Africa Cup of Nations (ACN) football matches, using the D4D dataset. For Ivory Coast's matches, the national team of the studied population, a small spike in the duration of calls is detected hours before the game, followed by an important drop as the game starts. For matches involving other national teams, a smaller reduction is mostly detected around games. By clustering base stations according to their behavior, using Ward's method, the authors identify regions far from the average behavior, mostly located around big cities or bordering to other footballing nations.

Activity patterns may also vary according to the characteristics of the special events. Xavier *et al.* [51] highlight such differences, as they detect an explosion in the number of calls on the New Year's Eve followed by a sharp decrease, unlike in the case of sport events.

Besides social events, natural hazards and disasters can have a significant impact on the phone usage patterns of customers. Pastor-Escuredo *et al.* [52] investigate the impact of fires taking place in Ivory Coast and observe that behaviors can vary across different regions: the following day after a fire takes place, they notice that, in rural areas, people make more calls in the morning than the evening, while typically it is the reverse; in small cities, they detect a large increase of terminating calls on that morning; while in big cities, they notice that fires result in a reduced calling activity. Elzen *et al.* [53], in turn, show that confrontations between political and ethnical groups lead to positive or negative variations in the mobile traffic activity, depending on the examined location with respect to the clashes region.

Other works focus instead on techniques to detect outlying behaviors. In an early study, Candia *et al.* [54] propose to detect anomalous events according to the gap between the observed number of calls and the mean number of calls over groups of geographically close base stations. They observe that the number and areas detected as anomalous vary with respect to the detection threshold. Dixon *et al.* [55] apply a similar methodology, as they detect anomalous events causing large variations in the number of calls over sets of base stations. They are able to identify events with increases in volume such as the New Year's day, and events implying decreases in the number of calls such as the birth of Prophet Mohammad, Easter Monday and January the 1st.

Altshuler *et al.* [58] deal with special events from a different perspective. They aim at detecting special social events by observing the communications mobile call graph. They propose a strategy in which they focus on the individual behavior of hub nodes, i.e. nodes with a high number of edges, to detect local outliers in the neighborhood of each one of them and finally identify the social events according to the prevalence of local outliers. Their results show that the proposed strategy is more efficient than tracking random users.

Key points. Special events are observed to imply special mobile traffic dynamics over the network. However, there is no clear view on how different special occasions would affect the evolution of traffic. Additionally, several methods have been proposed to detect outlying traffic behaviors. Nevertheless, we still lack a standard method, relying on a clear understanding of traffic dynamics, that allows to do so.

3.4 Individual traffic characterization

Analyzing individual mobile traffic behaviors complements studies of the evolution of aggregate traffic, by characterizing and understanding how individual customers consume network resources and use services. Previous works on the subject pay particular attention to the characterization of per-user activity, investigating its spatio-temporal variations and its correlation to the user's mobility.

3.4.1 Activity distributions

Individual consumption patterns can vary with respect to several aspects and can be very diverse. In an early study, Williamson *et al.* [26] analyze the calling behavior of 4,156 mobile users over one week and find that their usage patterns are very heterogeneous. Per-user activity is observed to follow a power law distribution, meaning that a vast majority of users perform only a few calls per week, while a notable amount of high-activity customers, generating hundreds of calls per week, still exist. Similar observations are also derived by Dasgupta *et al.* [72], as they analyze the calling traces of 3 million users over a longer duration of 1 month. Interestingly, they observe that the skewed distribution applies not only for voice calls, but also to data traffic. In turn, Paul *et al.* [28] affirm this conclusion at a larger scale spanning over a whole country. They show that high-end users can generate 100,000 times the median data traffic of all customers, which leads to having 10% of customers consuming 60% of network bandwidth. Shafiq *et al.* [29] also confirm the presence of a skewed distribution, as they notice that 5% of users are responsible for 90% of the total data traffic. Mobile customers are also heterogeneous in their access to specific services, as observed by Ben Abdesslem *et al.* [73]. In fact, they show that 20% of YouTube users are responsible for 78% of the total corresponding mobile devices requests.

Besides the differences among them in terms of consumed traffic, users can present diverse temporal consumption patterns. Candia *et al.* [54] detect that the per-user inter-call time follows a truncated power-law distribution, indicating that it is very rare to find consecutive calls made by the same user, separated by more than two months. However, Willkomm *et al.* [39] do not agree on this conclusion, and observe that inter-arrivals follow an exponential distribution.

In terms of activity durations, an agreement is observed over the fact that users tend to make short communications. Willkomm *et al.* [39], and Dasgupta *et al.* [72] observe a clear tendency of calls to be short, with a peak at around 1 minute. Similarly, Mucelli *et al.* [31] find that 80% of users in Mexico City are active for at most 4 hours per day, while less than 5% consume services for more than 10 hours per day.

Key points. Individuals are observed to be very heterogeneous in their global consumptions. However, the causes behind this heterogeneity have not been addressed. Moreover, while temporal characteristics have been considered by a few studies, spatio-temporal characteristics, as well as special behaviors remain to be explored.

3.4.2 Mobile user profiles

As remarked by studies in Sec. 3.4.1, individuals can be very diverse in their traffic consumption behaviors. Nevertheless, people with similar lifestyles and/or habits can be expected to present the same consumption characteristics. Lin *et al.* [70] perform an early study over a set of 600,000 customers. They cluster their calling behavior, using k-means algorithm, according to a set of usage features, such as the duration of calls, and duration of idle periods. They unveil major categories for which they devise adequate offers. Later, Becker *et al.* [71] group customers according to their individual calling patterns over a week with information about their hourly calling and texting loads, using also k-means algorithm. Focusing on the characteristics of two main clusters, out of seven, the authors find out that one of them corresponds to the behavior of commuters with a high level of mobility, while the other one maps well to the behavior of students.

Cerinsek *et al.* [41] are instead able to separate behaviors into two typical categories: morning users and late evening users, by grouping profiles according to the daily and weekly activities. Mucelli *et al.* [31] group users according to their total data volume and number of sessions over a period of 2 weeks, using a combination of the UPGMA and *k*-means clustering techniques. Their strategy allows them to classify users independently over the two metrics. Based on the data volume, they differentiate between light and heavy users, while the number of sessions allows them to distinguish occasional users from frequent ones.

A different approach is adopted by Keralapura *et al.* [27], who aim at profiling 3G browsing activities. They propose a co-clustering strategy called Phantom, that runs over a bipartite graph linking users to their browsed URLs. Their technique reduces first the size of the graph by grouping similar URLs into the same category and employing a divisive hierarchical algorithm over the obtained graph. It then chooses a particular level of the dendrogram, implying the largest distance between a couple of clusters, to generate the required co-clusters of the initial URLs and users. The application of Phantom over a one day trace indicates the existence of some variability in the aggregate application usages over long time intervals of 6 hours, with a remarkable tendency for users to follow these aggregate behaviors over shorter timespans within these intervals.

Key points. Despite their heterogeneous behaviors, individuals still present common traffic characteristics. Accordingly, studies discussed in this section have shown that individual consumption patterns can be grouped into a few categories. However, the methodologies adopted are diverse. There is thus a need to define a standard strategy that allows to do so. Additionally, so far, the categories have been derived only according to temporal representations of consumption patterns, while exploring the geographical space can be beneficial for networking applications.

3.4.3 Traffic-mobility correlations

Several works aimed at investigating possible correlations between traffic consumption and user's mobility. In a seminal work, Williamson *et al.* [26] observe that there is no strong correlation between the level of user activity and cell site changes. However, this conclusion is derived based on a packet data call traffic dataset limited to 4,156 users, and thus does not necessarily hold for an entire population. Later, analyzing mobility events in a 2G network over a much larger datasets, Couronné et al. [74] show that a strong correlation exists between the number of locations visited by a user and the number of communication events he generates: a highly mobile person is observed to induce a greater probability to use a mobile phone. This result is also confirmed by Paul *et al.* [28], as they study data traffic consumption in a 3G nation-wide network with respect to mobility characteristics in terms of number of visited locations and the value of radius of gyration, i.e. a measure of the distance traveled by a user. They notice that the median traffic generated by a subscriber doubles as we switch from a customer with a low mobility to one with a high mobility.

Trestian *et al.* [45] complement these studies by considering a deeper investigation of the impact of mobility on the actual application consumption of users. They compute the level of correlation between the user's mobility and the kind of service he is accessing. They detect important differences. Streaming music is mostly demanded by users when they are stationary. This demand sharply decreases as user's mobility increases. On the contrary, email shows a strong positive correlation with mobility, i.e., it is accessed more and more frequently as subscribers become more mobile. Other applications, such as social networking, show instead maximum access probability in presence of moderate mobility.

Finally, while studying the temporal evolution of calls, Candia *et al.* [54] observe that the fraction of mobile users who are traveling and making calls simultaneously remains stable over time.

Key points. Overall, these works have pointed to existing correlations between the traffic demand and the general mobility level of users. However, more complete investigations, taking into account mobility patterns over space and time remain to be investigated.

3.4.4 Devices and applications

Traffic consumption also varies according to different types of mobile devices available on the market, due to the diversity in their computational and storage capabilities. Moreover, traffic consumption varies with respect to the type of applications that they run. Also, these equipments constitute only a part of the devices that can access the cellular network. In the context of emerging machine-to-machine (M2M) communications, an enormous number of intelligent devices, e.g. communicating vehicles and smart metering devices, can be expected to communicate through the cellular network with diverse traffic patterns.

Shafiq *et al.* [29] were the first to analyze the load induced by different types of mobile phone devices and applications. Their results show that mobile phone devices of different types generate dissimilar traffic. Additionally, they detect some diversity in usage patterns among devices of the same type, due to the variation in user behaviors. More generally, they observe that the distribution of network traffic is highly skewed: 5% of devices account for 90% of the total network traffic and 10% of applications are responsible for 99% of the flows. Later, Shafiq *et* *al.* [4] complement their earlier study by comparing the traffic resulting from M2M devices to that of smartphones. Overall, they observe that M2M devices generate less traffic than smartphones, but imply a larger uplink traffic than downlink. Different temporal traffic profiles are detected for different M2M categories, which led them to consider a clustering strategy to group similar traffic time series. More precisely, they run Ward's method over wavelet transforms of times series, i.e. a decomposition of time series into a series of sines and cosines. Based on the clustering strategy, M2M devices are grouped into two major categories. The first one presents a diurnal behavior that maps to the working and non-working hours, while the second one reflects a flat consumption shape over the whole day. When comparing the session characteristics of M2M devices to those of smartphones, they observe that smartphones are generally active for much more time than M2M devices, presenting higher session lengths, and smaller session inter-arrival times. Nevertheless, they also point at important variability in different categories of M2M devices.

As mentioned earlier, mobile traffic can vary according to the application generating it. Zhang *et al.* [30] analyze mobile data traffic and find that applications providing similar services can in fact yield quite heterogeneous packet inter-arrivals. They thus identify sub-categories of social, news, and video applications that show comparable packet, flow and session-related metrics, by using *k*-means clustering algorithm and applying Principal Component Analysis [75], which allows them to understand the impact of each metric.

Key points. These studies have pointed out the variability in traffic patterns implied by different devices and applications. The derived findings remain to be verified over different datasets.

3.5 Conclusion

In summary, previous works, focused on the characterization of mobile traffic datasets, have brought out major properties concerning the dynamics of network traffic. Analyses of the aggregate behavior unveiled regular spatio-temporal patterns of mobile traffic, but pointed as well to some variability over time and space at particular granularities, due to land use of geographical areas and/or general societal lifestyles. Exceptional events are observed to induce irregular traffic behaviors, which vary depending on the nature of the event. Finally, in terms of individual consumption patterns, a significant heterogeneity is noted at a per-user, -device and -application levels.

Overall, networking analyses of mobile traffic datasets have revealed important traffic evolution features. However, the derived properties remain general, in the sense that they are not sufficient to derive proper models capable of capturing detailed spatio-temporal evolution of traffic, while this is particularly helpful for building common scenarios for testing and evaluating networking solutions. Clearly, such a step requires a deep understanding and characterization of the traffic evolution, which we lack today. Additionally, to acquire such a comprehensive knowledge of consumption patterns, adequate techniques and analysis tools need to be designed, an important aspect which has not received the necessary attention so far.

Chapter 4

Spatio-temporal cellular network usage profiling framework: the D4D case

4.1 Introduction

As discussed in Chapter 3, previous works analyzing mobile traffic datasets have disclosed general properties concerning the evolution of mobile traffic in cellular networks. However, they have failed to capture a clear view of its underlying fundamental properties, essential to derive proper representative usage profiles, that can simultaneously reflect aggregate and detailed views of traffic dynamics. From a networking perspective, this is particularly important for the design of dynamic cellular network management solutions, that automatically select among a set of system configurations, the most adequate one.

Potential applications include adaptive spatio-temporal resource control techniques, that aim at improving the performance of cellular networks, based on a limited set of system configurations, so as to avoid complex frequent reconfigurations. If we take the example of solutions aiming at offloading hotspot areas, then one would need to define a clear set of typical and particular usage behaviors, that drive the design guidelines of different configurations targeting specific spatio-temporal activity hotspots.

In this chapter, we introduce a mobile traffic datasets characterization framework, capable of processing very large datasets and automatically outlining network-wide utilization patterns. The particularity of this framework resides in the fact that it combines a set of models and methodologies, allowing to reflect a comprehensive view of traffic dynamics.

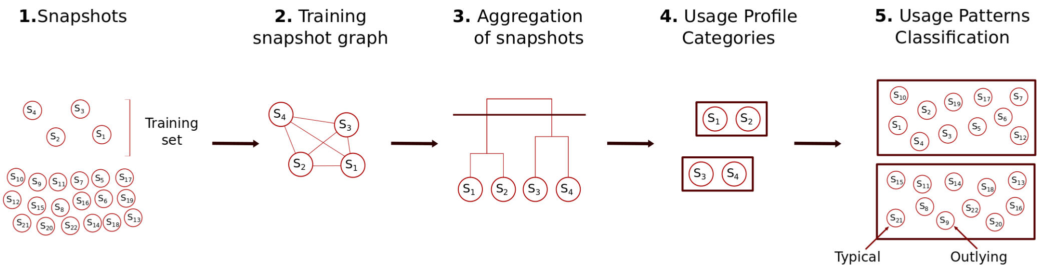

1.Snapshots
2. Training snapshot graph
3. Aggregation of snapshots
4. Usage Profile Categories
5. Usage Patterns Classification
Training set
Typical
Outlying

FIGURE 4.1: General workflow of the framework for the definition of categories of network usage profiles and their classification.

First, the framework operates over a set of snapshots of mobile traffic demand, which provide a description of usages over different geographical areas at separate time intervals. This definition of a snapshot allows to capture a complete view of the aggregate traffic, while maintaining a picture of simultaneous more localized usages. Instead, previous studies have widely employed time series [4, 11, 19, 20, 26–42, 44–55], which are capable of representing traffic evolution over one specific spatial granularity.

Second, two complementary distance measures are introduced in the framework to compare pairs of snapshots. These metrics measure the level of dissimilarity between traffic patterns, in terms of volume variations and volume distributions over different geographical areas, an important aspect that has not been considered in previous works. In fact, the majority of related works, tracking the evolution of traffic, are able to account only for volume variations separately for individual areas and lose the link between macroscopic and microscopic variations [4, 11, 19, 20, 26–55]. A similar observation holds for the few exceptions in the field, considering the distribution of traffic, whose scope, additionally, remains limited only to visual representations of normalized traffic volumes over geographical areas [33, 36].

Third, we define traffic patterns categories, by grouping similar snapshots together, according to the two distance metrics, based on a method combining a set of major data mining techniques. This allows to uncover hidden aspects of spatio-temporal traffic evolution, which have not been captured in previous mobile traffic analyses [4, 11, 19, 20, 26–55, 58, 59].

Fourth, we employ the generated categories as a basis for an additional traffic pattern classification step, allowing us to tell apart typical from outlying behaviors. The originality in this step resides in the fact that it operates over pre-defined categories of mobile traffic patterns accounting for both traffic volume and distribution variations among diverse geographical areas; while similar studies have been limited to simple strategies that rely on evaluating gaps in aggregate traffic volume, with respect to a certain defined normal behavior [54, 55].

In the rest of the chapter, we describe the different components and operations of our framework, whose general workflow is depicted in Fig. 4.1. We also present the results of its evaluation over the D4D CDR datasets.

Variable	Significance
i	Snapshot
k	Cluster
t	Observation time interval
z	Geographical area
$\overline{c}_k^n$	Center of cluster $\mathbb{C}_k^n$
$\overline{c}_k^n$	Centroid snapshot of all snapshots in the training set $\mathbb{S}'$
d_{ij}	Distance between two snapshots i and j, representing $\mathcal{V}_{ij}$ or $\mathcal{D}_{ij}$
e_{ij}	Edge in the training snapshot graph, linking snapshot i and snapshot j
w_{ij}	Weight assigned to edge e_{ij} in the training snapshot graph
v_i^z	Mobile traffic volume in snapshot i over the geographical area z
C^n	C index at level n of the clustering algorithm
C	Simplified notation of C^n
G	Training snapshot graph
CH^n	Calinski and Harabasz index at level n of the clustering algorithm
DH^n	Duda and Hart index at level n of the clustering algorithm
DH_{crit}^n	Critical value of the Duda and Hart index
B^n	Measure of clusters separation
F^n	F-ratio calculated over clusters at an iteration n of the clustering algorithm
F_{crit}^n	Critical value of the F-ratio
P_k^n	Measure of the proximity of snapshots in the same cluster
P^n	Measure of the proximity of snapshots in the same cluster for all clusters
P_{min}^n	Sum of the S^n smallest pairwise distances over the training set
P_{max}^n	Sum of the S^n largest pairwise distances over the training set
S^n	Total number of pairs of snapshots belonging to the same cluster
V_i	Total traffic volume recorded in the whole studied region over snapshot i
$\mathcal{D}_{ij}$	Traffic distribution distance between snapshots i and j
$\mathcal{V}_{ij}$	Traffic volume distance between snapshots i and j
$\mathbb{C}_k^n$	A cluster k at an iteration n of the clustering algorithm
$\mathbb{E}$	Set of edges in the training snapshot graph
$\mathbb{S}$	Set of snapshots
$\mathbb{S}'$	Training snapshot set
$\mathbb{T}$	Set of observation time intervals
$\mathbb{Z}$	Set of areas
$\mathbf{C}^n$	Set of clusters obtained at iteration n of the clustering algorithm
α	Standard normal score

TABLE 4.1: Main variables employed throughout the chapter.

4.2 System modeling

Our framework runs over a set of snapshots of mobile traffic demand. We define a snapshot as a representation of the global traffic load generated by users on the access network, at a certain time interval, over a set of geographical areas. This generic definition provides the framework with a significant level of flexibility over three dimensions: time, space, and traffic type, which can be adapted according to the objective of the study. In fact, a snapshot can describe voice, texting or data traffic, over a short time interval, e.g. of several seconds, or a longer one, e.g.

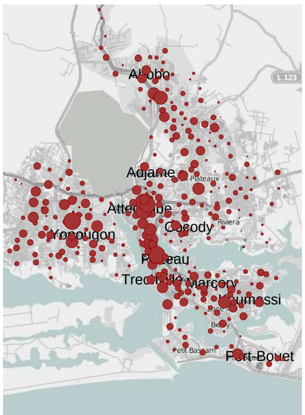


Figure 4.2: An example of a snapshot representing calling activity of Orange customers in Abidjan on April 12th, 2012 at 10:00.

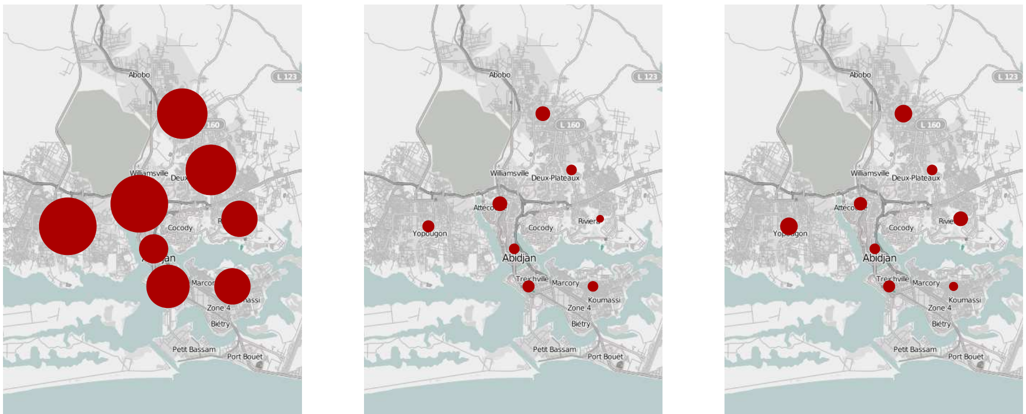
of a few hours; with information obtained at individual base stations or aggregated over larger geographical areas. Nevertheless, the finest levels of granularity, that can be employed, remain determined by the original dataset. We show in Fig. 4.2 an example of a snapshot representing with circles the total volume of calls over each base station in Abidjan, between 10:00 and 11:00, on April 12th, 2012.

In the following, we denote as $\mathbb{T}$ the set of observation time intervals, and $\mathbb{Z}$ the set of areas, over which the traffic volumes are aggregated. We use i to refer to a snapshot representing the demand during a certain time interval $t \in \mathbb{T}$. We employ $\mathbb{S}$ to refer to the set of snapshots over the whole observation period. Accordingly, v_i^z is used to indicate the mobile traffic volume observed in snapshot $i \in \mathbb{S}$ over the geographical area $z \in \mathbb{Z}$.

Our framework largely relies on the comparison of couples of snapshots, and more precisely, the similarity (or dissimilarity) between them. Indeed, different measures of similarity (or dissimilarity) can result in different outputs of the framework. In our study, we introduce two different measures that can convey different perspectives on the comparison between snapshots. In the following, we describe these measures and explain what each one represents.

4.2.1 Traffic volume distance

A first aspect to consider when comparing two snapshots, is the difference between them in terms of actual volume variations. As discussed in Chapter 3, many previous works focused on the evolution of the total volume on the access network, and were able to capture variations over time in their global behavior. Others considered aggregated traffic over a group of base stations or looked at each base station independently, and were able to derive spatio-temporal variations at finer granularities. Overall, these works are not able to capture a global view of fine-grained variations, and are able to either cover the total macroscopic evolution of traffic or its individual



(a) Snapshot A

(b) Snapshot B

(c) Snapshot C

FIGURE 4.3: Illustrative snapshot examples to convey the significance of the traffic volume distance $\mathcal{V}$.

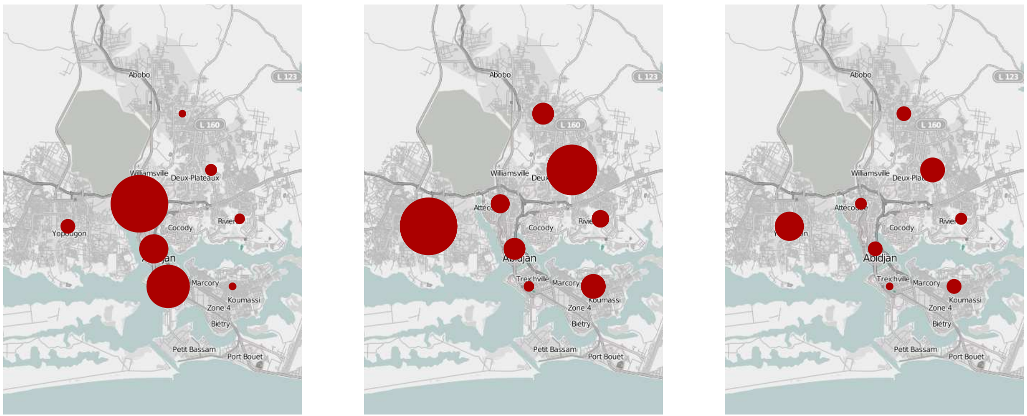
microscopic components, but not both. In our work, we fill this gap by introducing the traffic volume distance $\mathcal{V}_{ij}$, defined as follows, between a couple of snapshots i and $j \in \mathbb{S}$:

$$\mathcal{V}_{ij} = \sqrt{\sum_{z \in \mathbb{Z}} (v_i^z - v_j^z)^2} \tag{4.1}$$

Based on this definition, the traffic volume distance allows to capture the differences in actual volumes over distinct geographical areas between a couple of snapshots, through one single metric. If we consider that we have only one area in $\mathbb{Z}$, then $\mathcal{V}$ represents the total volume variation. If we divide the region of interest into a significant number of areas, then $\mathcal{V}$ can capture a more precise spatial diversity. We present in Fig. 4.3 toy snapshot examples, with the traffic per region z represented with a disk, whose size maps to the traffic volume. Comparing Snapshot A to Snapshot B would yield a high value of the traffic volume distance due to the notable differences between these snapshots in terms of volume over several geographical areas; while comparing snapshot B to snapshot C would result in a much smaller value of $\mathcal{V}$, as the traffic volumes remain quite similar over the set of areas.

4.2.2 Traffic distribution distance

The $\mathcal{V}$ metric alone is not sufficient to reflect a complete picture of the differences between a couple of snapshots. While it accounts for absolute variations of mobile traffic over separate areas, it disregards how the traffic is distributed among them. We thus introduce a second metric, the traffic distribution distance $\mathcal{D}_{ij}$, defined as follows for a couple of snapshots i and $j \in \mathbb{S}$:



(a) (b) (c)

FIGURE 4.4: Illustrative snapshot examples to convey the significance of the traffic distribution distance $\mathcal{D}$.

$$\mathcal{D}_{ij} = \sqrt{\sum_{z \in \mathbb{Z}} \left(\frac{v_i^z}{V_i} - \frac{v_j^z}{V_j} \right)^2} \tag{4.2}$$

where

$$V_i = \sum_{z \in \mathbb{Z}} v_i^z \quad \forall i \in \mathbb{S} \tag{4.3}$$

represents the total traffic volume recorded in the whole studied region over snapshot i. Thus, for a couple of snapshots, $\mathcal{D}$ considers the normalized volume variations over each area $z \in \mathbb{Z}$, instead of the absolute one, allowing it to capture differences between them in terms of traffic distribution over the different regions, independently of its absolute volume.

Fig. 4.4 shows toy snapshot examples, that can illustrate the differences captured by $\mathcal{D}$. While Snapshot A and Snapshot B present similar total traffic volume, they would imply a high value of $\mathcal{D}$ because of the disparity in the way the traffic is distributed over the different regions. Instead, comparing Snapshot B to Snapshot C would lead to a small value of $\mathcal{D}$, because of the similarity in the way the traffic is distributed, while the total traffic volume is clearly different.

4.2.3 Complementarity of $\mathcal{V}$ and $\mathcal{D}$

A natural question that one would ask is whether the two metrics $\mathcal{V}$ and $\mathcal{D}$ are needed for the comparison of real-world snapshots, or whether a single metric holds the same information that can be obtained based on the other one. To investigate that, we derive, first, the distributions of $\mathcal{V}$ and $\mathcal{D}$, over all couples of snapshots, in the D4D dataset. In Fig. 4.5(a) and Fig. 4.5(b), we show the corresponding Probability Density Functions (PDF) and Cumulative Distribution Functions (CDF). We notice that both PDFs are right-skewed, meaning that, although with a low probability, important differences can be observed between some couples of snapshots, according to each metric. However, the two distributions are clearly different: $\mathcal{D}$ presents a single maximum, while $\mathcal{V}$ presents three local maxima.

We further investigate these differences, by considering the variation of $\mathcal{V}$ with respect to $\mathcal{D}$. To that end, we calculate Pearson's Product Moment Correlation Coefficient (PPMCC) [76] [1]over the two variables to measure the linear correlation between them. 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 there is only weak positive relationship between $\mathcal{V}$ and $\mathcal{D}$. This can be also observed in Fig. 4.6, where each point represents, for 5% of all pairs of snapshots, the value of $\mathcal{V}$ with respect to $\mathcal{D}$. The plot visually confirms the weak correlation between the two metrics.

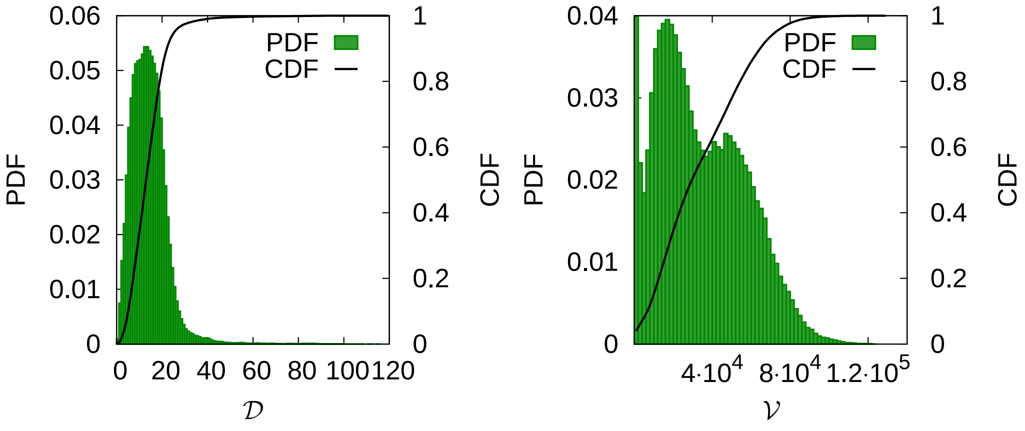
In fact, even for small values of $\mathcal{V}$, the volume distribution distance $\mathcal{D}$ can reach very high values, which reflect situations in which small volume variations over the set of areas would result in significant variations in the way the volume is distributed. Even more interesting is the case of small values of $\mathcal{D}$, for which $\mathcal{V}$ sweeps a wide interval. These represent snapshots with important traffic volume variations over the set of areas, which obey a common distribution of total traffic.

4.3 Defining usage profile categories

We define usage profile categories by performing the following three steps: *i*) Construction of a training snapshot graph, *ii*) Aggregation of snapshots, *iii*) Selection of network usage profile categories. We remark that the whole process is repeated twice, once for each metric. In the remainder of this section, we detail each one of these steps.

[1] For a pair of variables, with n samples, $x = (x_1, ..., x_i, ..., x_n)$ and $y = (y_1, ..., y_i, ..., y_n)$, the PPMCC is calculated as follows:

$$PPMCC(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad , \quad \bar{x} = \frac{\sum_i (x_i)}{n}$$



(a) (b)

Figure 4.5: (a): PDF and CDF of volume distribution distance $\mathcal{D}$ for all pair of snapshots. (b): PDF and CDF of volume variation distance $\mathcal{V}$ for all pair of snapshots.

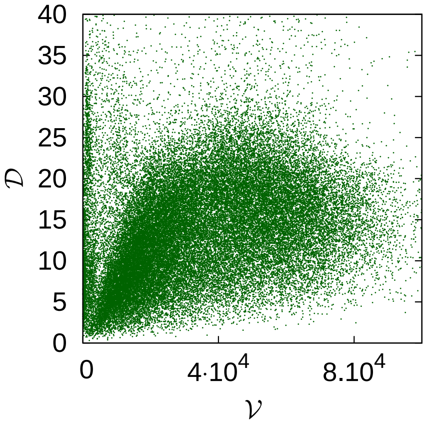


Figure 4.6: Scatterplot of distance measures $\mathcal{V}$ and $\mathcal{D}$ for each pair of snapshots i and $j \in \mathbb{S}$.

4.3.1 Training snapshot graph

We derive our usage profile categories over a training snapshot subset $\mathbb{S}' \subseteq \mathbb{S}$. The choice of $\mathbb{S}'$ has a direct impact on the quality of the obtained categories. The longer is the period that $\mathbb{S}'$ covers, the more representative are the derived categories. However, the complexity of the algorithms operating on top of the snapshots may constrain the size of $\mathbb{S}'$ for performance issues.

Once we extract snapshots in $\mathbb{S}'$, we map them to the vertices of an undirected weighted graph $G(\mathbb{S}', \mathbb{E})$, which we call training snapshot graph. The set of edges $\mathbb{E} = \{e_{ij} \mid i, j \in \mathbb{S}',\ i \neq j\}$ represents relationships between couples of snapshots in the training set $\mathbb{S}'$. The generated snapshot graph is in fact a clique as each pair of nodes shares an edge. We assign to each edge e_{ij} a weight $w_{ij} = \frac{1}{\mathcal{V}_{ij}}$, when focusing on the traffic volume metric and $w_{ij} = \frac{1}{\mathcal{D}_{ij}}$, when considering the traffic distribution metric. For simplicity, we employ, in the following, the notation $d_{i,j}$, to denote the distance between a couple of snapshots i and j, which can represent either $\mathcal{V}$ or $\mathcal{D}$.

4.3.2 Aggregation of snapshots

Once the training snapshot graph is obtained, we aim at finding network usage profiles categories by clustering the graph vertices. Given this objective, a first option would be to apply partitional algorithms which allow to group objects directly into a number of categories. However, two aspects would require careful investigations in this case. Partitional algorithms generally require a pre-knowledge of the number of categories to be formed, which, in our case, we do not possess. One may consider applying existing heuristics that can allow to infer that. Nevertheless, even in this case, the algorithm would require an initial affectation of objects per category, which are typically picked arbitrarily. Such a random choice can have an important impact on the way the categories are structured, and is thus not the best option.

A second possibility, that allows to cope with these aspects, is to consider a hierarchical clustering approach combined with a cluster selection strategy. In our case, we go for this option, and select the UPGMA algorithm, described in Sec. 3.2.2. We recall that the UPGMA relies on an agglomerative approach which starts from singleton clusters, including each, in our case, one node from the training snapshot graph. Then at every iteration, the algorithm merges the two clusters, or groups of snapshots, with the highest level of similarity, on average, according the distance measure $d_{i,j}$, and redefines accordingly the clusters for the following iteration. By the end of the execution of the algorithm, snapshots are organized in a dendrogram structure: a tree diagram outlining the partitionings that progressively group similar network usage profiles. To choose the categories, we study the dendrogram as detailed in the following step. We remark that we refer to the set of clusters at an iteration n, corresponding to the aggregation level n of the dendrogram, as $\mathbb{C}^n = \{\mathbb{C}_k^n\}$ where $\mathbb{C}_k^n \subseteq \mathbb{S}'$ denotes a cluster k at that level.

4.3.3 Selection of usage profile categories

In this phase, we explore the structure of the dendrogram generated based on the UPGMA algorithm 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 evaluated at each level of the dendrogram to quantify the separation among clusters. We referred to the extensive survey in [77] to choose the stopping rules. There, the authors compare the performance of 30 different indices in the literature, over a common dataset. We implemented and tested four different top-ranking indices according to the study: Calinski and Harabasz, Beale, Duda and Hart, and the C indices.

Ideally, for data organized in a clear clustered manner, with notable separations among each other and very cohesive internal structures, the four indices would point towards the same aggregation level reflecting the actual organization of objects. However, unless particular events take place, network traffic evolves smoothly, with no abrupt variations. This can result in some discrepancies among the results yield by the different indices, due to the diverse aspects that each considers. Thus, in case the different indices do not agree on the same aggregation level, we choose the one complying with most of them. Next, we describe the different indices.

4.3.3.1 Calinski and Harabasz index

For a generic level n of the dendrogram, the Calinski and Harabasz index, referred to as CH, is calculated as follows:

$$CH^n = \frac{B^n}{P^n} \cdot \frac{|\mathbb{S}'| - |\mathbb{C}^n|}{|\mathbb{C}^n| - 1}\ , \tag{4.4}$$

$$\text{with } B^n = \sum_{\mathbb{C}_k^n \in \mathbb{C}^n} |\mathbb{C}_k^n| \left(d_{\overline{c}_k^n, \overline{s}}\right)^2, \tag{4.5}$$

$$\text{and } P^n = \sum_{\mathbb{C}_k^n \in \mathbb{C}^n} \sum_{i \in \mathbb{C}_k^n} \left(d_{i, \overline{c}_k^n}\right)^2. \tag{4.6}$$

There, $\overline{c}_k^n$ is the center of cluster $\mathbb{C}_k^n$, a synthetic snapshot representing the centroid of snapshots in $\mathbb{C}_k^n$. It is obtained by averaging separately the traffic volume recorded over each geographical zone z, for all the snapshots of the cluster. Similarly, $\overline{s}$ is a synthetic snapshot, representing the centroid of all snapshots in the training set $\mathbb{S}' = \bigcup_{\mathbb{C}_k^n \in \mathbb{C}^n} \mathbb{C}_k^n$.

B^n measures 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 snapshots in the training set. Instead, P^n is a measure of the proximity of snapshots in the same cluster, it accounts for the distance between every snapshot i in a cluster $\mathbb{C}_k^n$ and the center of the cluster $\overline{c}_k^n$.

As n grows, B^n and P^n would respectively decrease and increase. The second factor in the CH expression compensates for that, as it becomes larger as the number of clusters $|\mathbb{C}^n|$ is reduced, allowing thus for a fair comparison between different aggregation levels.

Overall, the CH index compares the distance among clusters to their internal cohesion level. According to its definition, a better quality of clustering would be obtained for a higher value of the index. Therefore, the dendrogram level n, implying the highest value, is the one that grants the best separation among clusters, according to the CH index.

4.3.3.2 Beale index

The Beale index represents the F-ratio of a statistical F-test that accepts or rejects the hypothesis of merging two clusters at level n into a new cluster at level $n+1$. Suppose that, at 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:

$$F^n = \frac{P_m^{n+1} - (P_k^n + P_h^n)}{(P_k^n + P_l^n)} \Bigg/ \left(\frac{|\mathbb{C}_m^{n+1}| - 1}{|\mathbb{C}_m^{n+1}| - 2} \cdot 2^{2/|\mathbb{Z}|} - 1 \right), \tag{4.7}$$

$$\text{with } P_k^n = \sum_{i \in \mathbb{C}_k^n} \left(d_{i, \bar{c}_k^n} \right)^2 . \tag{4.8}$$

This F-ratio accounts for the variation of distance among snapshots within the two original clusters at level n and that among the same snapshots when they are grouped within the same cluster at level $n+1$. 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 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.

4.3.3.3 Duda and Hart index

Considering that, at level n, clusters $\mathbb{C}_k^n$ and $\mathbb{C}_h^n$ are grouped into cluster $\mathbb{C}_m^{n+1}$ at level $n+1$, the Duda and Hart index, referred to as DH, is derived as follows:

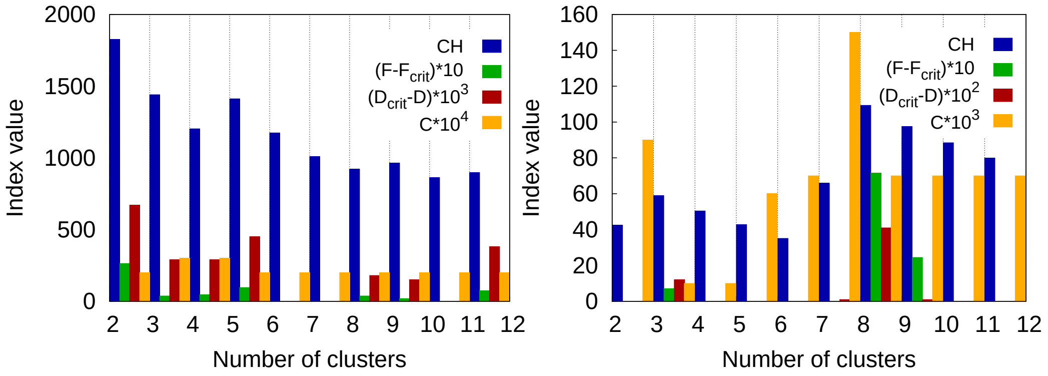
$$DH^n = \frac{P_k^n + P_l^n}{P_m^{n+1}}. \tag{4.9}$$

P_k^n is obtained according to the same expression as in the case of the Beale index. The DH index considers the ratio between the within cluster cohesion of the two separate clusters $\mathbb{C}_k^n$ and $\mathbb{C}_h^n$ and the level of cohesion when they are merged into the same cluster $\mathbb{C}_m^{n+1}$.

DH^n is compared to a critical value DH_{crit} derived as follows:

$$DH_{crit} = 1 - \frac{2}{\pi|\mathbb{Z}|} - \alpha \sqrt{\frac{2(1 - \frac{8}{\pi^2|\mathbb{Z}|})}{|\mathbb{C}_m^{n+1}|.|\mathbb{Z}|}}. \tag{4.10}$$

There, α is a standard normal score typically set to 3.2. The hypothesis of merging the two clusters is rejected if $DH^n \leq DH_{crit}^n$. In our case, we choose the best clustering quality, according to the DH index, as the one corresponding to a level n for which $DH^n - DH_{crit}^n$ is maximum.



(a) (b)

Figure 4.7: Clustering indices versus the number of clusters for a training set of two weeks according to: (a) the traffic volume distance measure $\mathcal{V}$, (b): the traffic distribution distance measure $\mathcal{D}$.

4.3.3.4 C index

The C index is calculated, for a level n of the dendrogram, as:

$$C^n = \frac{P^n - P^n_{min}}{P^n_{max} - P^n_{min}}.$$

P^n is derived the same way as in the case of the CH index. Accordingly, if S^n represents the total number of pairs of snapshots belonging to the same cluster obtained as:

$$S^n = \sum_{\mathbb{C}^n_k \in \mathbb{C}^n} \frac{|\mathbb{C}^n_k|(|\mathbb{C}^n_k| - 1)}{2},$$

then P^n_{min} is the sum of the S^n smallest pairwise distances, over the training set. Similarly, P^n_{max} is the sum of the S^n largest pairwise distances, over the training set. The best clustering quality is considered to be as the one corresponding to the smallest value of the index over all dendrogram levels: it would represent the farthest situation from the case in which we have the most dissimilar pairs of snapshots grouped in the same categories.

4.3.4 D4D usage profile categories

We derive network usage profile categories over the D4D dataset, by considering a two-week snapshot training set $\mathbb{S}' \subseteq \mathbb{S}$. We choose to train the framework on a two-week period for two reasons. First, this allows to cover the notable weekly periodicity of human activities. Second, this would reduce the impact of a possible bias that can be caused by a special behavior occurring during one of the two weeks, if any. Additionally, we generated the network usage profile categories over different random pairs of weeks from the five-month dataset. Interestingly, the number of obtained categories, as well as their content presented minor differences for the different pairs of training weeks. This validates our choice to consider a training set spanning over two weeks, as it leads to quite robust categories.

We show in Fig. 4.7 samples of the category selection process. The plots show the evolution of the four stopping indices introduced in Sec. 4.3.3, versus the number of clusters.

Focusing on the case of the traffic volume distance $\mathcal{V}$, in Fig. 4.7(a), we observe that all indices agree that the best separation between snapshots in $\mathbb{S}'$ is obtained for two clusters. Instead, for the traffic distribution distance $\mathcal{D}$, in Fig. 4.7(b), we notice that three, out of four indices, converge towards eight clusters, marking the best separation level among categories.

We present the structure of the categories found over a sample training dataset for $\mathcal{V}$ and $\mathcal{D}$ in Fig. 4.8(a) and Fig. 4.8(b), respectively. We notice that the two categories identified based on $\mathcal{V}$ clearly separate times with a relatively low activity, i.e., hours between 22:00 and 7:00, from 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 diverse areas of the city. The second category includes daytime snapshots from the weekdays, i.e., hours between 10:00 and 17:00, from Monday to Friday. These snapshots show a high concentration of mobile traffic activity in the office and university areas. The third major category contains most of snapshots of the weekend days, as well as early morning hours, between 5:00 and 9:00, and evening hours, between 19:00 and 22:00 of weekdays. The corresponding network usage is characterized by a high concentration of traffic in the residential areas. Also, five minor clusters appear, including a very small number of snapshots each.

4.4 Classification of usage patterns

Once the set of categories is identified over the training snapshot set $\mathbb{S}'$, we proceed to our classification step, which allows to cluster all traffic patterns from the whole observation period and tag their behavior as typical or outlying.

4.4.1 Extending usage patterns categories

In an iterative process, we assign all snapshots to a category, using the k-means algorithm, described in Sec. 3.2.2. We recall that k-means groups a set of objects into a pre-defined number

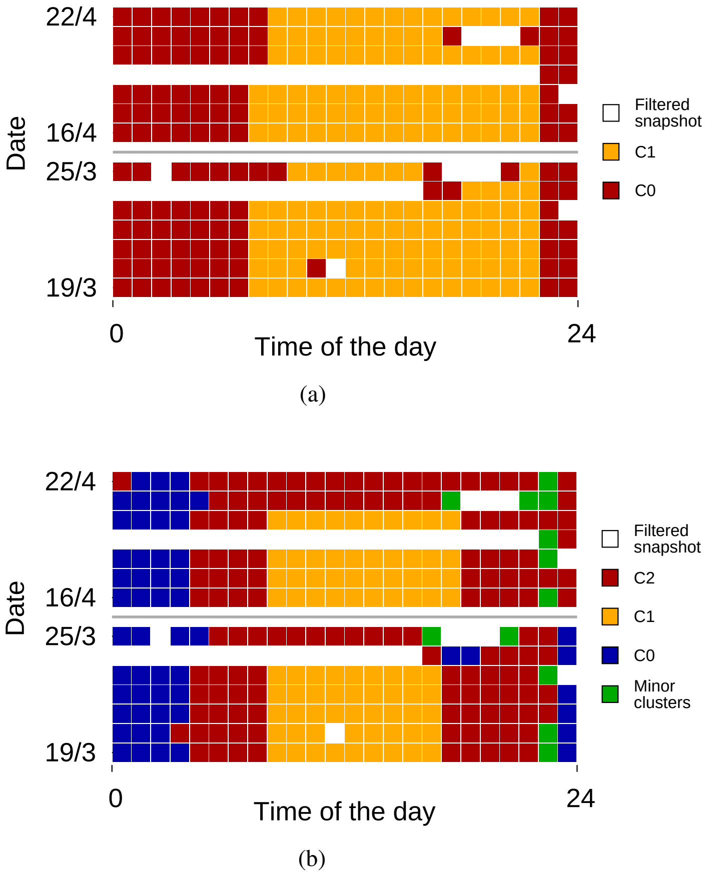
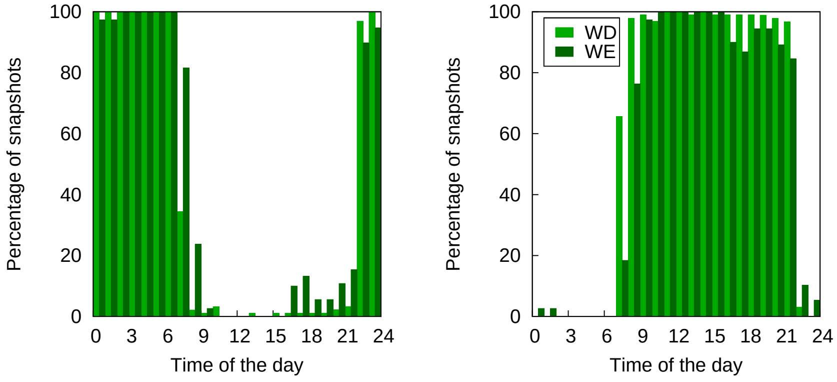


FIGURE 4.8: Content of mobile traffic profile categories defined on the training set $\mathbb{S}'$ composed of the two weeks March 19th–25th and April 16th–22nd, 2012, according to: (a) the traffic volume distance measure $\mathcal{V}$ and (b) the traffic distribution distance $\mathcal{D}$. Each square represents one snapshot, whose category maps to a color. White squares refer to snapshots that were filtered out from the dataset as discussed in Chapter 2.

of clusters k. Interestingly, in our case, the whole process of defining usage patterns categories provides a pertinent justification behind the choice of k. Even more, it lets k-means operate over an initial meaningful partitioning of the training snapshot set, instead of the typical arbitrary initial assignment of objects to clusters. We note that k-means relies on the distance between a snapshot i and the centroid of each cluster $\mathbb{C}_k^n$ in its decision. The algorithm assigns the snapshot to the category for which the measure is the smallest.

4.4.2 Labeling usage patterns behaviors

After assigning each snapshot to a category, we proceed to tagging the corresponding behavior. We do so, by exploring the way snapshots are placed into different clusters, as follows. For a particular time, of a certain day of the week, we assume that typical snapshots form the majority of snapshots and will be placed in the same cluster. Accordingly, by considering all snapshots



(a) C0: low activity category

(b) C1: high activity category

FIGURE 4.9: Classification of the five-month data using the distance measure $\mathcal{V}$. WD refers to weekday and WE refers to weekend.

TABLE 4.2: List of outlying snapshots, according to the classification provided by the measure $\mathcal{D}$.

Date	Assigned category	Typical category	Event
Sunday, Jan. 1st, 0:00	C2	C0	New Year's Eve
Saturday, Feb. 4th, 13:00	C1	C2	The Birth of the Prophet
Monday, Apr. 9th, 10:00 – 17:00	C2	C1	Easter Monday
Friday, Apr. 6th, 15:00 – 17:00	C2	C1	Good Friday
Wednesday, Dec. 7th, 18:00	C1	C2	Anniversary of the death of Felix Houphouet Boigny
Saturday, Jan. 7th, 11:00	C1	C2	Hilary Clinton and Kofi Annan's visit to Abidjan
Tuesday, Mar. 13th, 18:00	C4	C2	Election of National Assembly President and Prime Minister
Sunday, Dec. 11th, 19:00	C3	C2	New parliament election
Sunday, Feb. 12th, 23:00	C0	C2	Africa Cup of Nations final

occurring over the same repetitive weekly time interval, we check where the majority is placed and tag all corresponding snapshots as typical, while the rest, joining other clusters, would be labeled as outlying. As an example, if we have two clusters, with 95% of all Tuesday snapshots, at 10:00, joining the first one, then we would tag them all as typical, while the 5% in the second cluster would be labeled as outlying.

4.4.3 Classification of D4D usage patterns

We perform the classification step over the snapshots of the whole D4D dataset. We show the content of the different categories once the processing is completed. We plot in Fig. 4.9 and Fig. 4.10 the content of the categories defined with respect to the measures of $\mathcal{V}$ and $\mathcal{D}$, respectively. Therein, each plot describes the content of one category, by reporting the percentage of snapshots in $\mathbb{S}$, corresponding to a certain hour of the day and to either a weekday or weekend day, falling in that particular category.

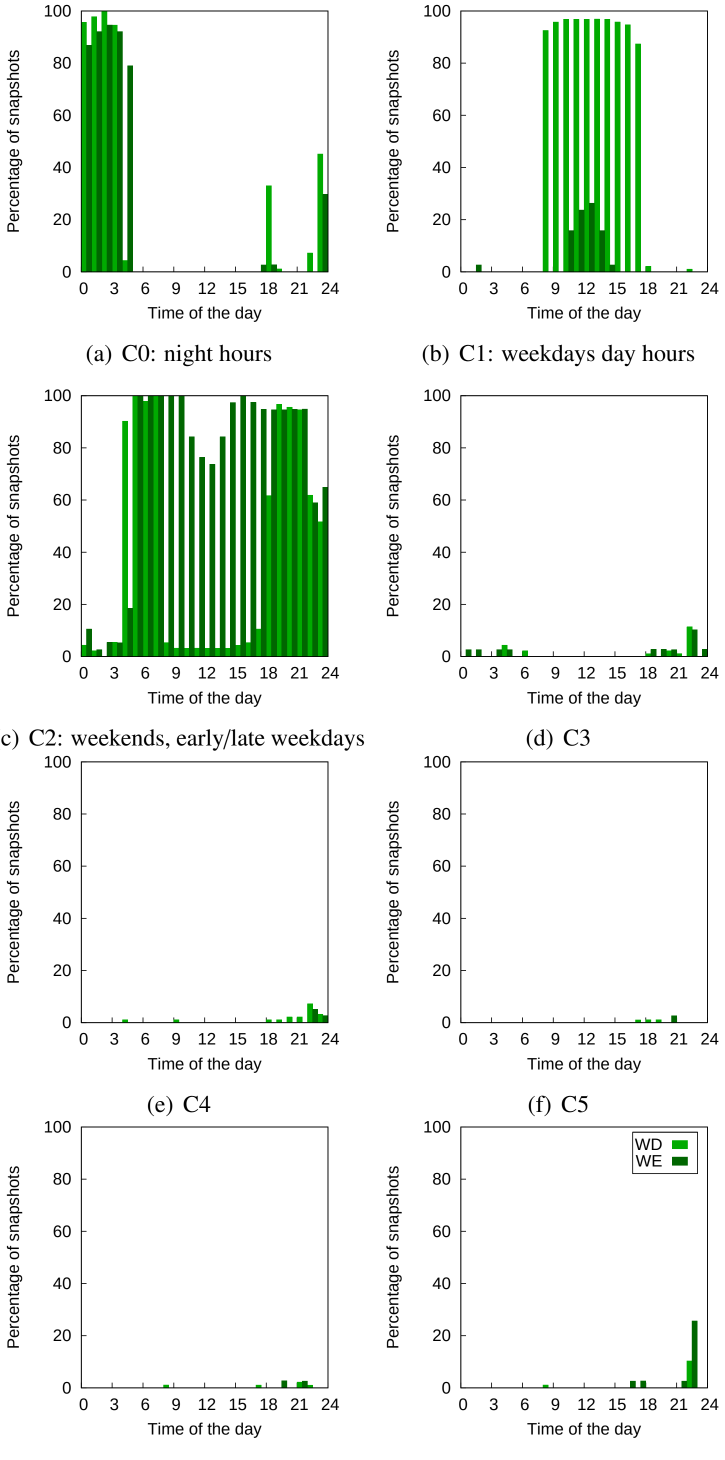
We can observe in both Fig. 4.9 and Fig. 4.10 that the categories preserve their initial structure, as most snapshots in the five-month dataset are classified in what we judge the corresponding typical category. However, we notice that some snapshots present outlying behaviors, as they join categories that differ from those they supposedly belong to.

Focusing on the two categories obtained based on the traffic volume distance measure $\mathcal{V}$, we observe that some snapshots at day time hours, such as 10:00 and 16:00, join the low-activity category in Fig. 4.9(a), while they would have been typically expected to be placed in the high-activity cluster. Clearly, these snapshots yield unusual network behaviors, in terms of number of calls. These outliers can be either due to technical problems in the network or electricity failures, despite our efforts to filter such snapshots as described in Chapter 2. As an example, we consider the case of Tuesday, March 20th at 10:00, whose calling activity is shown in Fig. 4.11(a), where each circle maps to one base station in the city, whose radius is proportional to the number of calls recorded in the corresponding cell over the one-hour snapshot duration. Comparing this snapshot to the one on Tuesday, April 3rd at 10:00, in Fig. 4.11(b), presenting a typical behavior at the same day of the week and the same hour, we can notice that an important number of antennas is missing and those remaining record only a minor traffic. This result demonstrates that the framework can identify unusual network behaviors with a higher accuracy than a simpler aggregate data analysis, which was already used in the filtering process and did not remove the snapshot in Fig. 4.11(a).

Concerning the outliers placed in the second cluster in Fig. 4.9(b), these represent low-activity hours showing an uncommon increase in mobile traffic and that are mostly related to special events. This it the case for example of the New Year's Eve, shown in Fig. 4.11(c), where clearly people are making much more calls than on a typical Sunday at the same time, January 8th, in Fig. 4.11(d). We also detected other outliers occurring on the Christmas' Eve, the day of the quarter final and the final football games of the Africa Cup of Nations.

For the $\mathcal{D}$ measure in Fig. 4.10, we can also notice multiple situations where a snapshot diverges from the expected behavior. In some cases, snapshots that would be typically placed in one of the major categories C0, C1, and C2 join another major category. This is the case, for example, of Friday, April 6th at 15:00, portrayed in Fig. 4.11(e), with respect to a typical behavior in Fig. 4.11(f), representing Friday, April 20th at 15:00. This outlying behavior happens to be the Good Friday, when the afternoon was a public holiday. This explains the reason why the snapshot is classified together with weekend snapshots in C2: it shows an increase in call volume in residential areas (Yopougon, Adjame, and Abobo), and a volume decrease in the largest office and commercial area of the city (Plateau). Similar observations hold for the other outliers in the C2 category, such as the whole day of Easter Monday, April 9th.

Similarly, outliers falling in the weekday daytime category C1 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 C0 reflect a reduced level of calls recorded in various areas of the city.



(a) C0: night hours

(b) C1: weekdays day hours

(c) C2: weekends, early/late weekdays

(d) C3

(e) C4

(f) C5

(g) C6

(h) C7

Figure 4.10: Classification of the 5-month data using the similarity measure $\mathcal{D}$. The three major categories C0, C1 and C2 are tagged with matching typical hours.

Other snapshots diverge from the typical behavior by joining minor clusters. These snapshots are found to be related to special events that do not concern the whole population of Abidjan and whose impact remains geographically localized. The corresponding distributions differ from what can be observed in a major category. This is the case for the snapshot on Sunday, December 11th at 19:00, in Fig. 4.11(g), occurring on the day of the election of a new parliament.

Comparing it to the typical corresponding behavior, such as the one on Sunday, January 22nd at 19:00, in Fig. 4.11(h), we notice that it is detected due to the increase of the calling volume in major residential areas (Yopougon and Koumassi), affecting the typical distribution of calls.

In Tab. 4.2, we present a list of the outliers detected according to the classification obtained with the $\mathcal{D}$ measure and whose cause we were able to identify. The table also shows the category where each outlier is placed, the category where it would have typically been placed, and the social reason behind the unusual behavior.

Another important observation concerning the categories obtained according to the $\mathcal{D}$ measure, 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 in Fig. 4.11(i) and Fig. 4.11(j) both belong to the C0 cluster, showing that, although people are making more calls on Saturday, December 24th at 23:00, the increase of activity is uniformly distributed over the entire city. 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.

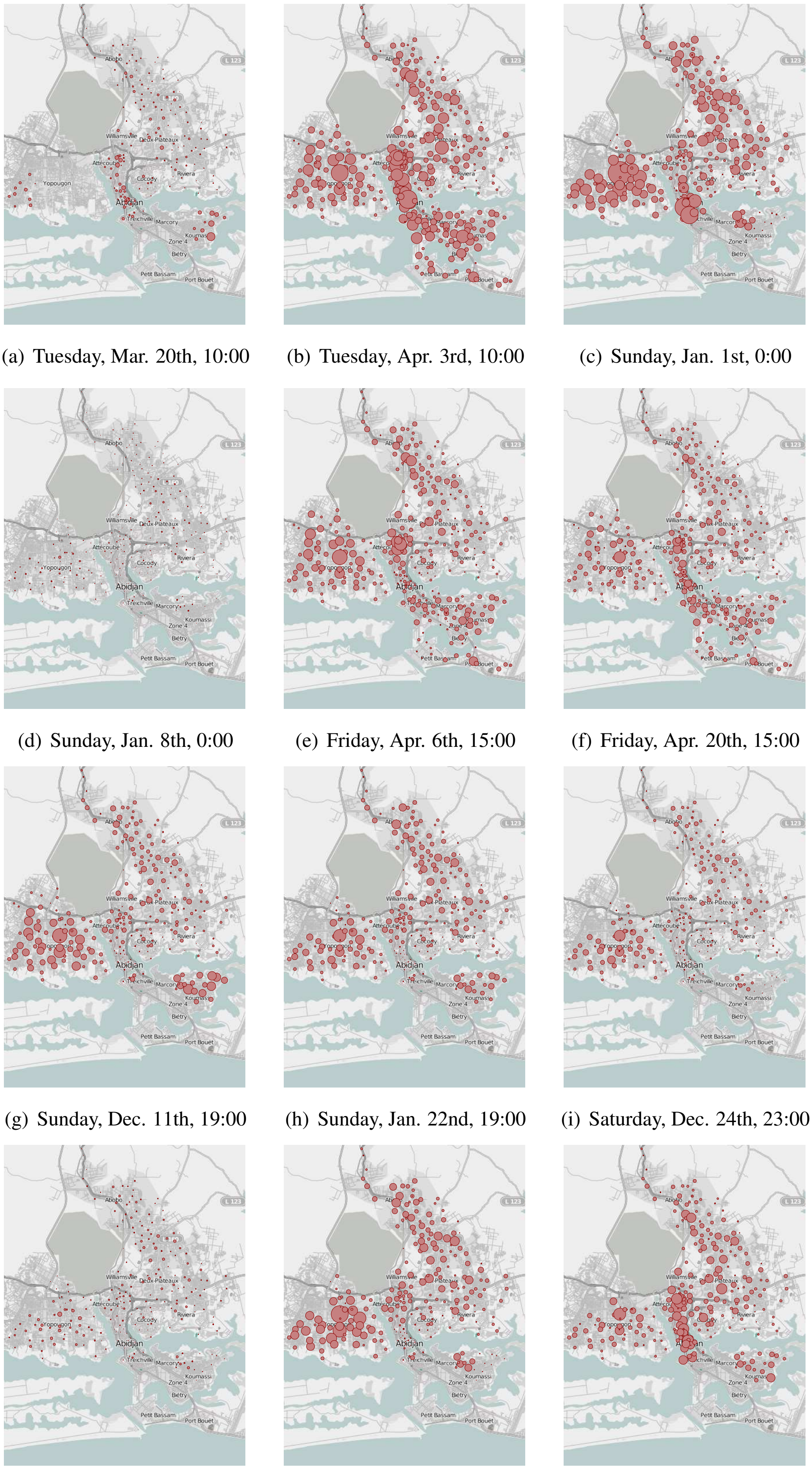
Finally, we remark that we also detected situations where snapshots were belonging to the same category based on $\mathcal{V}$, but were classified in different categories based on $\mathcal{D}$. This means that, for similar levels of volume, one can observe several volume distributions. Such it the case of the snapshots appearing in Fig. 4.11(k) and Fig. 4.11(l).

4.5 Conclusion

In this chapter, we presented a framework to characterize network-wide mobile traffic profiles. The framework builds categories of mobile traffic usages, by grouping together similar patterns from a training dataset. Similarity is evaluated according to a couple of distance measures that allow to capture complementary facets of mobile traffic profiles, at both microscopic and macroscopic levels. The categories are generated by combining a set of data mining tools in an original approach, and are used as a basis for a further classification procedure, over a complete very large dataset. This classification process allows to pinpoint outlying traffic behaviors corresponding to uncommon usages.

We apply the framework over the large-scale CDR D4D dataset covering the urban region of Abidjan. We started by demonstrating the significance and complementarity of the two distance measures, underlining the need for both to provide a comprehensive description of traffic dynamics. Then, by running our framework over the whole dataset, we showed how it coherently clusters mobile traffic patterns, according to the distance measures employed. Interestingly, we were able to map the obtained categories to meaningful social properties. At the end, we identified a set of unexpected behaviors in the mobile demand, which we relate to real-world events.

Finally, we remark that although in this chapter we apply the framework over GSM voice calling activities, its operations are generic and can be be employed over diverse kinds of traffic and contexts. In fact, it has already been applied over a Telecom Italia dataset with information about users demand in terms of calling, texting, and data traffic in Milan. The results obtained there underline again the capabilities of the framework to form significant categories of traffic.



(a) Tuesday, Mar. 20th, 10:00 (b) Tuesday, Apr. 3rd, 10:00 (c) Sunday, Jan. 1st, 0:00

(d) Sunday, Jan. 8th, 0:00 (e) Friday, Apr. 6th, 15:00 (f) Friday, Apr. 20th, 15:00

(g) Sunday, Dec. 11th, 19:00 (h) Sunday, Jan. 22nd, 19:00 (i) Saturday, Dec. 24th, 23:00

(j) Saturday, Dec. 17th, 23:00 (k) Sunday, Dec. 25th, 19:00 (l) Monday, Feb. 6th, 09:00

FIGURE 4.11: 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.

Part III

Networking Solutions and Mobile Traffic Datasets

Chapter 5

Technological exploitation of mobile traffic datasets

5.1 Introduction

Cellular network systems have long been designed and evaluated independently of the real-system dynamics. This is particularly true for research efforts in that direction, and is largely due to the fact that, in the past, we had no knowledge of how network usage patterns evolve over time and space.

Concerning marketing solutions targeting operator's products, it is evident that more detailed analysis of usage patterns can lead to more profitable techniques, that are beneficial for both the operator and the customers.

As for networking solutions, a wide range of infrastructure-oriented protocols and solutions have been proposed and examined over simplistic scenarios, that are not representative of real-world environments, and which can imply important biases in the evaluation results. The same is true for networking strategies targeting customers. The availability of mobile traffic datasets is changing the paradigm, by bringing performance evaluation environments a step closer to real world systems. More interestingly, they are paving the way towards more efficient usage-aware solutions.

In this chapter, we review previous works employing mobile traffic datasets to design and evaluate innovative solutions relevant to cellular networks. We discuss these studies by separating between: *i*) marketing strategies and *ii*) networking solutions.

	Study		Dataset							Focus						
	Name	Date	Operator	Area	Time	Users	V	T	D	CT	SA	LS	CS	DC	UQ	PS
Marketing	Wei [22]	08/02	–	Southern Taiwan	4 months (2001)	114 K	✓	✓		✓						
	Dasgupta [72]	03/08	–	–	5 months (2007)	3.1 M	✓			✓						
	Lin [70]	10/07	–	Northern PRC	–	600 K	✓	✓		✓						
	Cardona [35]	12/14	–	European country	7 months (2011/12)	40 K			✓	✓						
	Belo [78]	07/13	–	European country	1 year (2008/09)	10 K	✓	✓			✓					
	Szabo [79]	11/06	–	–	14 months (2004/05)	5.5 M	✓	✓	✓		✓					
Networking	Zang [12]	09/07	3G	3 cities	1 month (2006)	2 M	✓	✓	✓			✓				
	Xu [80]	06/11	–	LA, USA	1 week (2010)	–			✓			✓				
	Gerber [81]	03/11	–	USA	2 days (2010)	~ M			✓				✓			
	Finamore [82]	12/13	–	European metropolis	1 day (2012)	> 200 K			✓				✓			
	Shafiq [48]	06/13	–	2 metropolis	Several days (2012)	100 K	✓		✓					✓	✓	
	Wang [83]	05/09	–	–	6 months	100 K	✓	✓						✓		
	Agarwal [84]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓						✓		
	Zhu [13]	04/09	–	US	2 weeks (2008)	2 M		✓	✓					✓		
	Zhu [125]	05/13	Orange	Ivory Coast	2 weeks (2011)	500 K	✓	✓						✓		
	Yu [86]	04/13	–	Metropolis, PRC	1 month (2011)	65 K			✓						✓	
	Balachandran [87]	09/14	–	Metropolis, USA	1 month (2012)	1 M			✓						✓	
	Shafiq [88]	06/14	–	USA	1 month (2012)	500 K			✓		✓				✓	
	Zang [89]	09/11	–	50 states, USA	3 months (2010)	25 M	✓									✓
	Montjoye [90]	03/13	–	Western country	15 months (2006/07)	1.5 M	✓	✓								✓
	Song [91]	07/14	–	–	1 week	630 K	✓									✓
	Acs [92]	08/14	Orange	Paris, France	1 week (2007)	2 M	✓	✓								✓

TABLE 5.1: Main features of works employing mobile traffic data towards designing technological solutions. In the study columns, date is in MM/YY format. In the dataset columns, V is voice, T is texting, D is data. In the focus columns, CT is churning and traffic plans, SA is service adoption, LS is localization strategies, CS is caching strategies, DC is D2D communications, UQ is User QoE, PS is privacy solutions.

Tab. 5.1 summarizes the works presented in this chapter. It highlights the major features of the datasets they employ, in terms of originating operator, spatio-temporal span, number of concerned users and recorded traffic type. Similarly to mobile traffic analyses, discussed in Chapter 3, we notice here that the datasets remain also very diverse. Tab. 5.1 additionally indicates the network-related aspects considered in each study, which we discuss in detail in the chapter, providing a quick synthetized reference to the reader.

5.2 Marketing strategies

Mobile traffic datasets constitute a valuable source of information for the design of marketing strategies. Their analysis allows the operator to understand the behavior of customers, their calling patterns and habits, and thus to propose suitable business solutions and offers.

5.2.1 Churning and pricing strategies

The sector of cellular networks is mostly driven by highly competitive network operators on the market. With the diversity of proposed offers and pricing schemes, mobile users tend to change operators over time. Each operator remains interested in seeking strategies to attract customers and avoid the churning behavior, i.e. customers switching from one operator to another. Analyzing mobile data traffic can be a very useful step towards this objective, as it can help understand and anticipate the behavior of users, and in particular that of churners.

Wei *et al.* [22] propose a methodology that classifies users into churning/non-churning categories, strategically designed to account for the typical low percentage of churners. Their technique builds upon the features of individual calling behavior, such as the volume and frequency of calls. The evaluation of the methodology over a dataset of 114,000 customers indicates that it allows to correctly predict 70% of churners with a 20% false positive ratio.

Dasgupta *et al.* [72] also focus on the prediction of churners by taking into account the impact of social relationships among customers on their possible future churning behavior. More precisely, they expect that a customer's choice of switching to a new operator is influenced by the choice of his contacts. Accordingly, they propose a diffusion model operating over the users mobile call graph, which allows to predict future churners by examining social connections among customers. The diffusion process is observed to successfully predict 60% of future churners over the studied graph.

Clearly, predicting churners alone is not sufficient to limit their manifestation and increase the market penetration of the network operator. It needs to be complemented by well-designed traffic plans that prevent possible churners from letting go their current operators and also attract other operator's customers. Lin *et al.* [70] focus on this particular objective and group 600,000 customers into different categories, according to their common calling pattern characteristics. They then introduce new offers adapted to each group's calling features. The real-world adoption of the new offers by the concerned operator confirms the efficiency of the proposed strategy, with an increase of 64% in the number of customers.

In a similar perspective, Cardona *et al.* [35] investigate the profits of different pricing plans to customers, in terms of cost savings. They observe that collaborative pricing plans can be very beneficial for customers, with savings of up to 45% for pre-defined group plans, that allow users to share a common allowed capacity. Open sharing plans, providing customers with individual permitted traffic capacity that one can freely use and sell, can be even more beneficial, with savings that can reach 70%, with respect to baseline cost. Tethering schemes are also shown to imply additional gains, especially in dense urban areas.

Key points. Several techniques have been proposed to anticipate and reduce the churning behavior of customers. They have been tested separately over different datasets. Thus, their performance needs to be assessed over a common dataset that allows to compare one to the other. Additionally, the proposed techniques do not take into account offers proposed by other operators on the market, while being a factor of important influence on the churning behavior.

5.2.2 Service adoption

Belo and Ferreira [78] study the impact of the mobile call graph structure on the diffusion of telecommunication-related products among customers. They identify different adoption incentives, with respect to various products. As an example, they notice that an offer allowing users to make calls for free to all service adopters spreads wider than an offer that permits a user to communicate for free, with all the customers of the operator. They also observe that the diffusion mechanism can be positively or negatively influenced by social relationships.

Szabo and Barabasi [79] consider a similar problem and aim at evaluating the impact of the social relationships on the adoption of services. For social networking services, they observe a strong correlation between the user's adoption decision and the choices of his contacts. Instead, no correlation is detected for professional-oriented services, such as email and browsing.

Key points. Social relationships are observed to have an important impact on the diffusion process of services among customers. However, other factors affect the decision of a customer regarding the adoption of a service, such as his consumption needs and consumption restrictions imposed by operators. Accordingly, the impact of social relationships remains to be assessed in light of other factors.

5.3 Networking solutions

Novel practical solutions aiming at improving the performance of cellular networks have been proposed based on mobile traffic data analysis. They cover various technical aspects that target the performance of networking strategies, users' experience and, more generally, the whole system's efficiency.

5.3.1 Localization strategies

Locating users and events occurring over the network, with a high level of precision, by employing typical procedures, can be very costly for network operators in terms of signaling overhead. Intelligent localization schemes can be powerful tools in that sense. Zang and Bolot [12] precisely aim at improving the efficiency of the paging procedure, by adapting it to per-user mobility history, rather than covering large location areas with hundreds of cells. The proposed strategy is observed to imply a 90% reduction of signaling load, with respect to the typical paging procedure, at the cost of a slight increase in paging delay of less than 10%. Instead, Xu *et al.* [80] focus on the localization of IP-level measurements. They introduce a system called AccuLoc, which allows to map events accurately, once trained with precise network measurements. Their results show that their system is capable of mapping measurements to clusters of 4 cells with an accuracy of 70%.

Key points. Intelligent localization schemes are observed to significantly reduce signalling overhead in the network. While these strategies are beneficial from a networking perspective, they require saving measurements information from the history, with costs in terms of memory that have not been accounted for by these studies.

5.3.2 Caching strategies

Content caching strategies can bring notable benefits to the network from an infrastructure and a user perspectives. Gerber *et al.* [81] explore the potential of such mechanisms, by considering content caching, at different levels of the cellular network architecture. They notice that cache hit ratios of 33% can be obtained, by assuming an illimited caching capacity at the national data center. Additionally, by accounting for caching costs, in terms of processing, storage and transfer costs, caching is found to be best implemented at regional data centers with savings that can go up to 27%.

Finamore *et al.* [82] also focus on content caching, but consider a push strategy, according to which the content in the cellular network is pre-staged to the mobile device cache before it is requested. The authors evaluate three different caching strategies, accounting for the content popularity, its volume, and both simultaneously. They conclude that the content popularity-based strategy, which balances between the complexity of operations and cost savings, can lead to a reduction of up to 20% of downloaded volume, for a cache size of 100 MB, in case any popular content can be cached.

Key points. Content caching strategies are beneficial for the network. Proposed schemes consider caching at various levels of the network architecture. However, they have not explored correlations between users consumptions and their mobility, while this can lead to more important savings by considering schemes that move the content according to users mobility.

5.3.3 D2D communications

Mobile data traffic hold very powerful information for exploring the potential of emerging device-to-device (D2D) communications, which allow users to exchange data directly without going through the cellular infrastructure.

Wang *et al.* [83] were the first to investigate the capabilities of D2D communications, with a particular attention to the spread of malwares over the network. To that end, they consider an epidemic model to describe the transmission of malwares, such that a device can be either infected, i.e. carrying the malware and capable of transmitting it, or susceptible, i.e. being at risk of becoming infected. They remark that, although at low speed, D2D bluetooth communications allow the malware to hit all susceptible devices. In contrast to that, they consider the dissemination of the malware through multimedia messages, periodically sent to an infected user's contact every two minutes. They find the malware to spread much faster with multimedia messages than bluetooth contacts, while remaining limited by the market share of the particular operating systems. Similarly, Agarwal *et al.* [84] infer the potential of D2D communications in disseminating information at nation-wide scales, based on a more detailed information dissemination scheme, with additional latent infected and susceptible states, that consider situations in which the mobile device is switched-off. Therefore, the authors show that one single device can propagate information to 90% of a 5,000-user population, spread all over the Ivory Coast.

A complementary study to these works is completed by Zhu *et al.* [13], who propose to limit the spread of messaging viruses, by taking advantage of the structure of the mobile call graph among customers. More precisely, they suggest to split the graph into partitions and inject security patches to target users, staying at the border between different partitions. They consider two graph partitioning strategies. The first one, referred to as balanced graph partitioning, aims at forming even partitions by accounting for both the number of connections per user and the weights of these connections. The second one, called clustered graph partitioning, favorizes forming partitions with high internal cohesions, in terms of connection weights. By comparing these strategies to a random one, they observe that the clustered patching achieves the best performance, as it limits the infection rate within a reasonable bound much faster than the other two strategies. More precisely, they notice that if they begin patching after 2% of mobile devices had been infected, clustered partitioning bounds the infection rate to 0.025 within 30 time units.

Besides information dissemination applications, the capabilities of D2D communications have been considered to offload the RAN. Zhu *et al.* [125] investigate the possibility of opportunistically routing information among users. They evaluate the performance of six different opportunistic routing methods. Their results indicate that all algorithms perform well in dense small areas with active users, posing opportunistic communications as a promising solution towards delivering delay-tolerant data packets among cellular network users. Shafiq *et al.* [48] alternatively focus on radio access techniques. They propose to take advantage of D2D communications to share a single connection to the RAN. They show that their approach performs particularly well in crowded events, leading to 95% less failed connections.

Key points. D2D communications can facilitate the propagation of information over the network. Their performance is constrained by the employed technology, the will of users and application requirements, a few elements that require further investigations.

5.3.4 User QoE

In the field of cellular networks, efforts are mostly concentrated on the performance of the infrastructure. Not much attention has been paid to the overall system efficiency, and in particular, user experience is disregarded although being of high importance. The following works consider several aspects relating to the quality of experience (QoE) perceived on the user's side.

Yu *et al.* [86] focus on the energy efficiency of transitions among different 3G user equipment radio resource connection states, allowing to control each user's access to network radio resources. Their analysis confirms that a significant amount of power is wasted, during inactivity times, as the user switches among the various states. Based on that, the authors investigate the temporal correlations of traffic workload and propose a prediction model of future data transmissions, in order to effectively cut unnecessary waiting times. Their results indicate that, on average, 56% of energy can be saved, when applying their scheme. Shafiq *et al.* [48] also investigate the impact of mobile device radio resource connection state transitions on the QoE perceived by users. Focusing precisely on a crowded event context, with local overload of the network, they show that a slight decrease in the value of state changing timeouts, of one or two seconds, can result in a better tradeoff between usage of radio resources, energy consumption and delay.

Previously discussed studies focus on user experience independently of the considered networking contexts. Works by Balachandran *et al.* [87] and Shafiq *et al.* [88] rather target performance metrics on the user side for specific applications. Balachandran *et al.* [87] study HTTP web browsing sessions and radio-level signalization to understand the impact of technical network metrics, such as handovers, failures, power levels and throughput, on the mobile user browsing experience, in terms of incomplete downloads, abandoned sessions, and session length. They unveil a set of major parameters, allowing to fully characterize and accurately anticipate the QoE of users. These include inter-radio-access-technology handovers, the energy per chip of the pilot channel over the noise power density, i.e. a measure allowing to quantify how well a signal can be distinguished from the noise in a cell, and the load in the cell. Surprisingly, they realize that this set does not include factors that are often considered important by network operators, such as, the average radio link data rate, soft handovers and inter-frequency handovers.

Shafiq *et al.* [88] concentrate alternatively on video streaming applications to mobile users. They analyze HTTP records and user connection measurement reports, and study the impact of technical network parameters on the video abandonment. Based on their analysis, they provide operators with guidlines allowing to improve users QoE, relating to video streaming. As an example, they note that a 1-dB increase of signal-to-interference ratio reduces the video abandonment probability by 2%. Additionally, the authors introduce a methodology to determine whether a user will complete a streaming video session, or not, depending on radio network statistics and information extracted from TCP/IP headers. The model is capable of predicting the complete download of a video by a mobile user, with an accuracy of 87%, by observing only the initial 10 seconds of a session.

Key points. It is possible to predict user's QoE with a high accuracy. Improvements of QoE have been considered based on the radio resource connection state transitions. However, improvements of QoE with respect to individual applications have not been explored yet.

5.3.5 Privacy solutions

While forming a very rich source of information for several fields of study, mobile data traffic still raises concerns about the privacy of users. In most of the previous works, mobile operators anonymize customers identifiers, so as to preserve their privacy. However, such measures may not be sufficient, due to the critical information that mobile data traffic can infer. A few datasets provided a higher level of anonymity by decreasing the spatio-temporal granularity of datasets. Still, the applied granularities were not chosen according to clearly defined rules, which motivated investigations of the granularities allowing to uniquely identify users and triggered efforts targeting the design of adequate anonymization techniques.

Zang *et al.* [89] inspect the number of per-user locations needed to uniquely identify a user. Towards this end, they evaluate the anonymity of datasets, based on the number of users sharing the same top-visited locations. Their results show that, at cell and per-event granularities, knowing the top three locations is sufficient to uniquely identify more than 50% of users. Furthermore, the authors study how varying the spatial granularity of the shared data, by aggregating it over geographical areas, can affect the level of anonymity. They notice that city-wide aggregations are required for efficient anonymization. They also consider a temporal domain approach, according to which they renew users identifiers periodically. They find such a strategy to ensure privacy for daily updates.

These results were later confirmed by Montjoye *et al.* [90], by considering a random choice of per-user spatio-temporal points, instead of most visited locations. Their results indicate that if the location of an individual is specified on an hourly basis, and with a cell spatial granularity, four randomly chosen points are enough to uniquely characterize 90% of the users, whereas two randomly chosen points still uniquely characterize more than 50% of the users. The authors thus explore the impact of a reduced spatio-temporal granularity on the anonymity of dataset, but find that even coarse resolutions may provide little anonymity, stressing the importance of carefully choosing the adequate resolution of shared mobile data traffic.

Similarly, Song *et al.* [91] evaluate the uniqueness of trajectories, over a one-week dataset, which represents the probability of finding a couple of trajectories, including the same set of points. Their results show that more than 60% of trajectories can be uniquely identified with only two randomly selected points. This percentage grows up to 95% with a set of four random points. Thus, they test a temporal domain approach with identifiers updated every six hours, which leads to a value of 40%, for the case of two randomly selected points. The benefit remains less important for a higher number of points.

Finally, Acs *et al.* [92] focus instead on preserving user anonymity, when releasing aggregate spatio-temporal density information. Their main concern relates to areas with low counts, that risk to unveil individual mobility patterns. To handle that, they introduce a scheme which combines sampling, clustering and filtering processes of per-cell information, to generate aggregate density information for areas grouping several cells. Their strategy is shown to provide a good tradeoff between privacy guarantees and data utility level.

Key points. There is an agreement on the fact that sharing several spatio-temporal points can threaten users privacy. A number of strategies based on varying the spatio-temporal granularity of shared datasets have been proposed. However, we still lack a clear view of the adequate granularities to consider. Further studies over different datasets are required.

5.4 Conclusion

In this chapter, we reviewed previous studies that introduce and evaluate technological solutions relating to cellular systems, using mobile traffic datasets. Overall, novel ideas, tackling important networking and marketing problems, have been proposed and complemented by realistic evaluations. However, most of the introduced strategies are static, meaning that they are adapted to the general mobile traffic features. Exceptional works consider dynamic techniques that follow the evolution of traffic. In fact, for networking applications, a very promising research direction is the conception of dynamic strategies that explore the spatio-temporal variability of mobile traffic demand. There a variety of topics remain to be explored including network planning and management solutions.

Chapter 6

Assessment of achievable energy savings in real-world cellular networks

6.1 Introduction

The exponential increase of mobile traffic, over the last few years, has pushed the network community to seek solutions that allow to accommodate the increase of demand. A variety of strategies have been proposed. These solutions aim at providing higher capacity by *i*) targeting the network infrastructure through the densification of the network with additional base stations (BSs), *ii*) offering a better utilization of the frequency spectrum through cognitive radio techniques, or *iii*) offloading traffic through wifi and device-to-device communications.

Network operators have widely resorted to the network densification option, which has led to a significant increase in the number of base stations deployed in the world. In France, there are currently more than 100 K base stations, operating across the various 2G, 3G and 4G standards [93]. While such an operation allows to solve the capacity problem, it remains inefficient from an energy consumption perspective, raising environmental concerns on one hand, and amplifying the expenses on the network operator side, with the growing fuel prices, on the other hand.

The problem comes from the fact that BSs, dimensioned with respect to the peak traffic hours, are always switched on and consume a high amount of energy at all times, while traffic demand can undergo important fluctuations. We can observe such a significant variability in Fig. 6.1, for a representative working day, in the city of Abidjan. The figure captures the evolution of the total number of hourly calls throughout the day and shows that important differences can be detected between day and night. Even more, the figure tracks the evolution of the total number of hourly occupied GSM resources, representing allocated time-frequency blocks over GSM frames. This curve is derived according to the total duration of hourly calls, that can reflect a

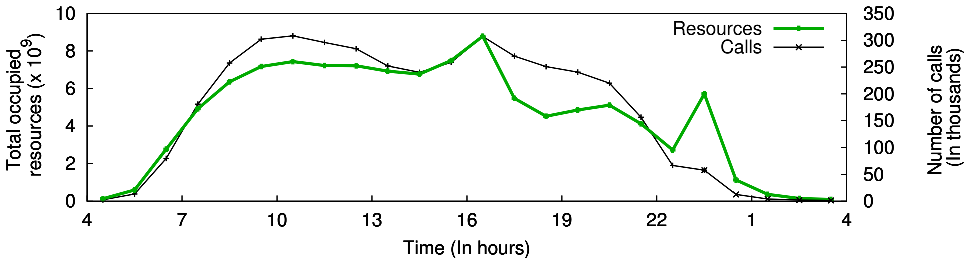


Figure 6.1: Total number of hourly occupied GSM resources and voice calls, over a typical working day in Abidjan.

more precise view of users activity over the network. Interestingly, we notice that the number of occupied resources presents more fluctuations than the number of calls, with a remarkable peak at 23:00. That is due to the fact that the duration of calls does not necessarily follow the evolution of number of calls. In particular, at 23:00, users make a low number of calls, which last for long times, translating into a high occupation of network resources.

As BSs are typically responsible for more than 50% of the total power of a wireless cellular network [94], considering strategies that allow to tune their energy consumption to the traffic demand, with adaptive power control mechanisms, clearly yields a high potential. If we take the example of a commercial center, we would need several BSs to cover its peak traffic demand. However, during the night, they would not all be required, in order to accommodate the minor traffic demand remaining in the area. Accordingly, a BS power control strategy, with switch-off possibility, over the set of BSs surrounding the commercial center, would be very beneficial from an energy consumption point of view. Additionally, for BSs with several antennas, one may consider such power control strategies over individual antennas. Although operating at the level of antennas leaves space for a finer adjustment of energy consumption, it would necessitate more complex and costly procedures to be implemented, especially when expanded to a city-wide of network-wide level.

We thus focus, in our work, on power control strategies over BSs. Yet, even there, one would need to capture a clear view of traffic dynamics. Mobile data traffic constitutes a very valuable source of information in this context. By providing a realistic evaluation environment that captures real-world usage patterns, it allows to assess possible real-world energy savings resulting from such load-adaptive network management strategies.

Based on this, we aim, in this chapter, at evaluating how much energy can be saved in the network with a dynamic switch-on/off strategy, complemented by flexible power transmission levels, in order to maintain a geographical coverage of the network and accommodate the demand of customers. More precisely, we address the following questions: *i*) Which is the base station configuration that allows to maintain a geographical coverage of the network? *ii*) When can such a configuration satisfy users demand? *iii*) How can we adapt this configuration to the traffic demand at other times of the day?

The rest of the chapter is organized as follows. Sec. 6.2 reviews major previous works dealing with power control mechanisms of base stations, so as to reduce network infrastructure energy consumption. Sec. 6.3 introduces some system parameters and assumptions, useful for our evaluations. Sec. 6.4, 6.5, 6.6 answer the previously introduced questions, by combining optimization and heuristic techniques. In Sec. 6.7, we discuss possible extensions of our model. Finally, Sec. 6.8 concludes this chapter.

6.2 Literature review

The topic of green cellular communications has received significant attention in the research community over the last decade. A wide range of solutions have been studied [95–97]. Proposed techniques can operate at various levels. Some strategies operate at the level of individual base stations, by improving the material efficiency, introducing power saving schemes, and using renewable energy resources. Other solutions aim at reducing energy consumption from a network planning perspective. There, the focus is on the benefits resulting from dense heterogeneous network deployments. Finally, a third category considers the system design, and aims at enabling green communication in cellular systems through novel technologies, such as cognitive radio and cooperative relaying.

Our work focuses on switch-on/off techniques with flexible cell sizes, and thus operates at individual base station level. When considering such operations, two constraints would naturally emerge: the geographical coverage over the zone of interest, and the accommodation of users demands. While all previous works dealing with switch-on/off strategies take into account the latter condition, a large majority of works disregards the former one and replaces it, instead, with the coverage of users appearing in the network, at a particular time [98–106]. Clearly, this would result in coverage holes, which would prevent potential users appearing there from connecting to the network.

In particular, early works [98–100, 104] aimed at assessing potential gains, that can result from switching off a certain fraction of BSs in the network at night hours, based on simplistic strategies and analytical models. Chiariviglio *et al.* [99] present a strategy that allows to switch off a predefined portion of base stations during a night interval, determined by balancing the switch-off period duration and cells radii, while meeting an access blocking probability limit. The proposed method starts from an initial configuration and iteratively swings between two operations: reducing night interval duration or increasing the power of BSs, until acceptable cells radii are obtained. Their results indicate that savings can go up to 50%. Micallef *et al.* [104]

assess the potential of the switch-on/off technique by proposing to switch off underutilized, randomly selected antennas or BSs. Their evaluations show that applying such a technique at an antenna-level leads to higher savings than a BS-level one, that can go up to 45%.

In a complementary work, Marsan *et al.* [98] introduce an analytical model to evaluate energy savings in the network, in case a fraction of BSs can be switched-off during low-traffic periods. They observe that energy savings of at least 25% can be achieved. Kelif *et al.* [100] also consider an analytical study that allows them to understand how a reduction in the transmission power leads to a decrease in the coverage of users and the global network capacity. Marsan *et al.* [101] extend the scope of these analyses to the level of two collaborative cellular access networks and estimate possible gains according to various strategies, that can balance different metrics such as switch-off frequencies, roaming costs, and energy savings on the two operators sides. Their results indicate that savings of at least 15% can be obtained.

Later, more flexible techniques that adapt to the evolution of traffic, and thus can include switched-off base stations at any time of the day, were proposed by Dufková *et al.* [102] and Oh *et al.* [103]. Dufková *et al.* [102] adopt a quite unique approach, in which they model the problem with a bipartite graph, where BSs and users are represented with nodes, while edges refer to associations among them. By modifying the structure of the graph based on the coverage and capacity constraints, they observe that energy savings can reach up to 50%. Alternatively, Oh *et al.* [103] propose a decentralized switch-on/off strategy, in which a BS coordinates with neighboring BSs and decides its configuration, based on the calculation of a network impact factor, reflecting the impact of the switch-on/off decision on the network, in terms of variation of the traffic load and service rate. Their results indicate that savings can reach around 50% during weekdays, and 80% during weekends.

Previously discussed works allow to infer the potential of switch-on/off techniques in terms of energy consumption, but consider that a BS can either be on, typically at its maximum transmission power, or off, and thus disregard possible intermediate power levels. Niu *et al.* [105] take into account this aspect and propose a centralized and a decentralized cell zooming and reassociation strategies that aim at minimizing the energy consumption in the network, while still minimizing the access blocking probability. Their results indicate the presence of a trade-off between energy consumption and outage probability for each algorithm. Liu *et al.* [106] also consider that base stations can be attributed diverse power states, but focus on a heterogeneous network context. The authors propose random and strategical sleeping policies, that aim at maximizing the energy efficiency, based on some predefined fractions of base stations in the various power states[1]. The main idea behind these algorithms is to define the state of the BS according to the number of users that can be associated to it, the distance between the BS and the UEs,

[1] The energy efficiency is defined as the average ratio between the network throughput and the power.

as well as some mobility information. The evaluation of their strategy implies that the energy efficiency can be improved by a percentage that varies between 15% and 30%.

However, all of these works focus on providing coverage to users in the network, and disregard the actual geographical coverage of space. Peng *et al.* [109] complete the only study that takes into consideration this constraint. They do so, by defining equivalent base stations that can replace each other, from a geographical coverage perspective, with an increase of power transmission level. They select from this set BSs that need to remain active to accommodate peak traffic, and those that need to remain active during low traffic hours. They then propose a procedure to switch on base stations as the traffic increases in the network over groups of equivalent base stations. While operating over the set of equivalent base stations allows to fulfill the geographical coverage constraint, it risks to operate over very reduced sets of equivalent base stations, including two or three BSs, and thus can lead to important additional unnecessary energy consumption with a significant amount of bordering regions that can be covered by several BSs.

Other studies fulfilled the geographical coverage constraint, by keeping all BSs on and tuning the usage of resources to users demands. Tipper *et al.* [107] consider adaptive service and frequency dimming schemes, according to which the authors propose to find the set of frequencies and services that are sufficient to carry all traffic demand at a minimum power cost. By evaluating the strategy over diverse scenarios, the authors observe that significant savings can be achieved. In a similar study, Combes *et al.* [108] modify the state of resources, defined as the number of transmitters for 2G systems and the number of carriers for 3G and 4G systems, while ensuring a Quality of Service (QoS) limit for users. Their simulation results show that savings of around 30% can be attained. While such strategies can lead to notable gains, it is evident that additional complete switch-off of base stations would be more beneficial.

Studies by Marsan *et al.* [111] and Han *et al.* [110] aimed at pushing energy savings on the electricity grid even farther, by considering that BSs can be supplied with renewable energy sources. Marsan *et al.* [111] propose to solely rely on renewable energy sources, for powering base stations, and observe that with sleep mode strategies, the number of photovoltaic panels can be reduced to its half, with respect to an always-on case. Alternatively, Han *et al.* [110] consider hybrid energy supplies, and aim at optimizing the coverage of each BS, so as to minimize the overall on-grid energy consumption in the system. By comparing their strategy to a best-effort approach, they observe that their algorithm allows to reduce the on-grid energy consumption.

In addition to disregarding the geographical coverage, previously discussed works tend to lack in realism, in one or several aspects, including network deployment, traffic profiles and scale of study. In fact, several studies, including recent ones, consider regular network deployments for their evaluations [98, 99, 105, 106, 110], which are far from real deployments. Moreover, a significant number of works remains limited to a small set of tens of base stations [98, 99, 102–105, 107, 108, 110] which can only be representative of one district of a large city. Besides, these studies largely rely on unrealistic traffic profiles, with trapezoidal [98], sinusoidal [99, 101] and unrealistically simulated traffic with no clear motivations for the chosen user density [100, 102, 105, 110]. Other studies, relying on real data, account for aggregate traffic patterns [103, 108]. Those considering real per-BS traffic remain an exception and deal only with normalized traffic, for proprietary reasons [109].

In our work, we aim at filling these gaps, by introducing power control strategies accounting for the two major coverage and capacity constraints. First, we derive a minimum power-consuming BSs configuration, that allows to satisfy the geographical coverage constraint. Second, we test at which times of the day this configuration allows to accommodate users demands. Third, we propose a technique that allows to adapt this configuration to higher traffic demands, at lowest increase in power consumption costs. Additionally, we evaluate our methodologies over a city-wide real-world network deployment with realistic per-BS traffic profiles.

6.3 System parameters and assumptions

In this section, we present several system parameters and assumptions, that we need to define, in order to assess the system's power consumption. We remark that Tab. 6.1 summarizes all variables employed in this chapter.

Users demands. We evaluate the proposed strategies, by testing them over snapshots of GSM per-frame users demands, obtained over one day and separated by ten-minutes time intervals. The per-frame demand is derived from real-world hourly per-BS traffic profiles, as detailed in Sec. 2.3.3.1. We note that our solutions are adapted to GSM systems, as we possessed mobile traffic datasets emerging from them at the moment of the study. Extensions to other cellular network systems are discussed in Sec. 6.7.

User-BS association. Given the per-frame demand, we consider that a user appearing in a certain area i of the city, gets associated to the closest BS j, from which he receives a signal stronger than -100 dBm. We note that each area i maps to a square of 100 m by 100 m of the geographical space, as employed in Sec. 2.3.3.1. Also, we evaluate the path loss between a BS j and a user in area i using Walfish-Ikegami propagation model [24], parameterized with the same values as in the case of demand generation in the same section.

BS capacity. The D4D dataset does not specify the capacity of BSs. Thus, for each BS, we define it according to its maximum per-frame demand, over the whole day. Clearly, this may not necessarily represent the actual deployment configuration, but one would expect operators to dimension the system according to some knowledge of users consumptions, providing more

Variable	Significance
a	Power model parameter: dynamic consumption power coefficient
b	Power model parameter: static consumption power
f_i	Demand of users in area i
i	area
j	BS
x_{ij}	Association variable between area i and BS j
y_{ij}	Number of radio resources allocated by BS j to serve users in area i
P^c_j	Power consumed by a BS j
$P^{c,max}_j$	Maximum power consumption of BS j
P^c_{ij}	Power consumed by a BS j in order to cover an area i
P^t_j	Power transmitted by a BS j
R_j	Number of radio resources for BS j
$SINR_{ij}$	Signal-to-interference-plus-noise ratio for a user in area i with respect to the signal received from BS j
W	Bandwidth for one LTE radio resource time-frequency block
$\mathcal{A}$	Set of areas
$\mathcal{A}_j$	Set of areas associated to BS j
$\mathcal{A}^{max}_j$	Set of areas covered by BS j when transmitting at maximum power
$\mathcal{A}^{vor}_j$	Set of areas belonging to the Voronoi cell of BS j
$\mathcal{B}$	Set of BSs
$\mathcal{B}_i$	Set of BSs that can cover area i
$\mathcal{B}^{max}_i$	Set of BSs that can cover area i when transmitting at maximum power
$\mathcal{B}^{on}_{geo}$	Set of switched-on BSs to provide the geographical coverage
$\mathcal{N}_j$	Set of neighboring BSs to BS j

TABLE 6.1: Variables employed throughout the chapter.

resources where higher activity levels typically emerge. We assume that a BS is formed by three sectors and consider three classes of low, medium, and high capacity BSs. In GSM, capacity is accounted for in terms of number of carrier frequencies. The smallest scheduling unit is the frame, including eight time slots, where each carrier frequency can carry users traffic and/or signaling information. Thus, a carrier frequency that only holds voice traffic can serve eight users during a frame. Generally, a BS reserves the first time slot of one of its carrier frequencies for signalling messages. During communications, it also passes some signalling information, instead of voice traffic, over the slots dedicated for voice traffic. For simplicity, we consider that, for each frame, a BS uses two slots to carry signalling traffic, regardless the number of used carrier frequencies. This leads us to the following distribution, with respect to the maximum per-frame demand over each BS: 207 BSs with a low capacity of 2 carrier frequencies (CF) per sector, 144 BSs with a medium capacity of 4CF/sector and 10 BSs with a high capacity of 6CF/sector.

Power consumption. We derive the power consumed by a BS j, P^c_j, based on its transmission power P^t_j, by applying the following model for macro BSs [112]:

	2 CF/sector	4 CF/sector	6 CF/sector
a	24.1	47.6	70.2
b	463.5	759.2	894.7

Table 6.2: Power consumption model parameters, for diverse BS resource capacity [112]. a is the coefficient that controls the dynamic power consumption, that varies with respect to the transmitted power and b represents the constant power consumption part relating to the electronic equipments.

$$P^c_j = \begin{cases} aP^t_j + b & \text{When } P^t_j > 0 \\ 0 & \text{Otherwise} \end{cases} \tag{6.1}$$

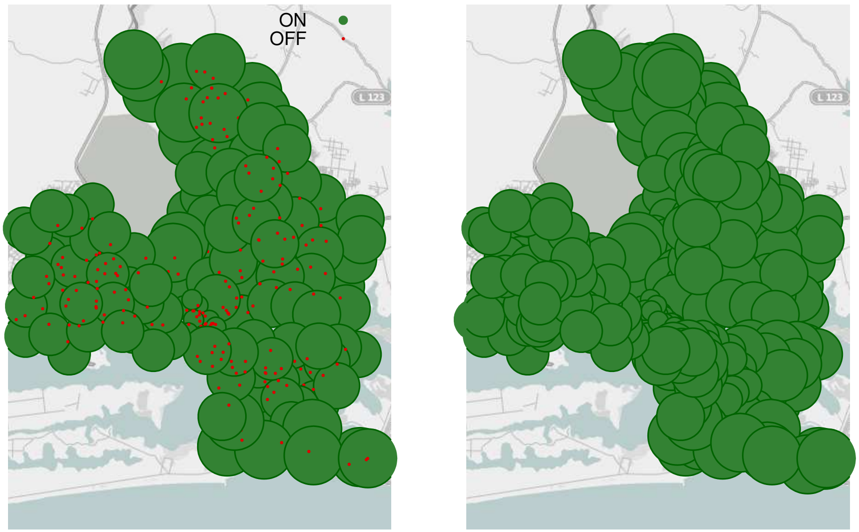
a is the coefficient that controls the dynamic power consumption, that varies with respect to the transmitted power and b represents the constant power consumption part due to the electronic equipments. The values of a and b are obtained by considering the capacity of each BS as specified in Tab. 6.2.

Additionally, we consider that for each capacity category, a BS transmission power is limited to a set of values, in compliance with GSM technical specifications [113]. In terms of consumed power, this means that P^c_j would also acquire a value from a limited set $\{0, P_1, ..., P_6 = P^{c,max}_j\}$. The maximum possible consumed power $P^{c,max}_j$ is evaluated with respect to a maximum transmission power of 20 W.

6.4 Geographical coverage access network configuration

As discussed in Sec. 6.2, the large majority of previous works aiming at reducing network infrastructure energy consumption, with power control mechanisms, have ignored the need to maintain a network coverage over the geographical space. Therefore, in this section, we derive a minimum power consuming BS configuration that allows to provide a full coverage of the geographical space. In this configuration, all BSs that are on, need to remain on at all times, and transmit at least at the minimum power suggested by the solution. We model our problem as an optimization problem as follows.

We define $\mathcal{A} = \{1, ..., m\}$ as the set of areas that should be covered in the studied region, and $\mathcal{B} = \{1, ..., n\}$ as the set of deployed BSs. P^c_{ij} is a constant representing the minimum required consumed power by BS j, that permits it to provide coverage to a user, at the farthest edge of area i. To derive the power consumption thresholds P^c_{ij}, we first evaluate the required transmission power P^t_{ij} of the BS j to serve area i, by applying the Walfish-Ikegami empirical propagation model [24], and then computing P^c_{ij} by applying the power consumption model in equation 6.1. We use x_{ij} to represent the association between area i and BS j, such that, x_{ij} is equal to 1 if area i is covered by BS j, and 0 otherwise. Additionally, we denote the areas that BS j can cover



(a)

(b)

Figure 6.2: Configuration of BSs to provide a full geographical coverage over the city with: (a) a switch-on/off scheme, where a BS can be on or off and (b) all-on scheme, where a BS cannot be switched-off. A switched-on BS is represented with a green disk, whose size is an approximation of its actual coverage area, and a switched-off BS is referred to with a red dot.

at maximum power as $\mathcal{A}_j^{max}$, and the set of BSs that can cover an area i when transmitting at maximum power as $\mathcal{B}_i^{max}$.

Our decision variables are the continuous variables, $P_j^c\,\forall j \in \mathcal{B}$, and the binary variables $x_{ij}\,\forall i \in \mathcal{A}_j^{max}, \forall j \in \mathcal{B}_i^{max}$. This problem needs to be solved only once, its variables are not time-dependent. In fact, by solving it, we derive the configuration of BSs that need to be on at all times, at the minimum consumption power determined in the solution.

$$\min \sum_{j\in\mathcal{B}} P_j^c$$

$$P_j^c \geq P_{ij}^c x_{ij} \quad \forall i \in \mathcal{A}_j^{max} \tag{6.2}$$

$$\sum_{j\in\mathcal{B}_i^{max}} x_{ij} \geq 1 \quad \forall i \in \mathcal{A} \tag{6.3}$$

Constraint (6.2) allows an area $i \in \mathcal{A}_j^{max}$ to be covered by a BS $j \in \mathcal{B}$, if the power consumed by the BS is higher than the constant P_{ij}^c. Constraint (6.3) imposes that each area should be covered by at least one BS.

We remark that $\mathcal{A}$ is determined by considering the subset of all areas in the city which can be served if all BSs are transmitting at maximum power. We attempted to solve our problem over the whole city at once, however, we were not able to obtain a solution in a reasonable time, due to the high complexity of the problem in terms of number of integer variables. We thus had to

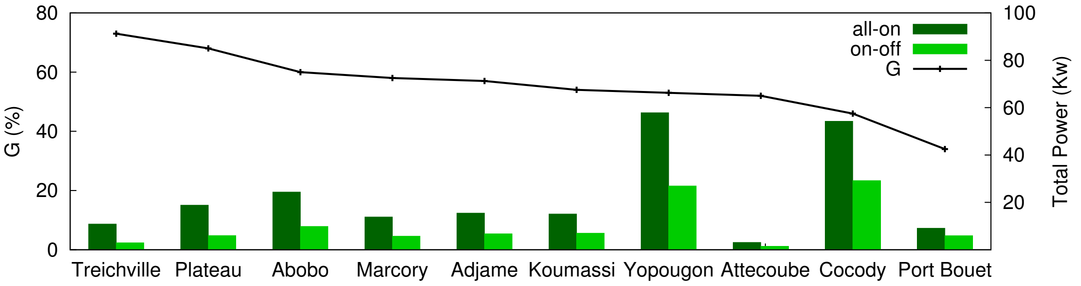


FIGURE 6.3: The obtained gain (G) when switch-on/off strategy is applied to define geographical coverage BS configuration, with respect to the all-on strategy. The figure also portrays the total power consumption per commune, for each case.

resort to smaller zones, and considered, instead, the different communes in Abidjan. Fig. 6.2(a) shows the configuration that we obtain. There, each green disk represents an approximation of the covered area by each BS that needs to remain on. Switched-off BSs are referred to with red dots. We notice that different communes can bear diverse fractions of BSs to be switched-off. As an example, the highly dense city center shows a significant portion of base stations that can be switched-off, with respect to other parts of the city. This is a consequence of the diversity of signal propagation characteristics in the different communes of the city, as discussed in Chapter 2, and the actual density of deployed BSs.

We compare the total consumed power with this switch-on/off scheme, to the one computed when all BSs are on, referred to as all-on. The latter is obtained by solving the same optimization problem, as before, with the additional constraint:

$$P_j^c > 0 \quad \forall j \in \mathcal{B}. \tag{6.4}$$

This constraint imposes that each BS is switched-on and transmits at least at minimum power. Fig. 6.2(b) portrays the corresponding configuration. By comparing it to the switch-on/off configuration in Fig. 6.2(a), we notice that they both provide the same geographical coverage. Fig. 6.3 shows the normalized gain G in the power consumption for each commune, together with the absolute power value for both strategies. The obtained gain over each commune varies from 34% to 73% indicating that important energy savings can be obtained.

In terms of power consumed per BS, we show, in Fig. 6.4(a) and Fig. 6.4(b), the distributions obtained with the switch-on/off and all-on strategies respectively. Overall, we can observe that up to 60% of base stations consume 0 W, with the switch-on/off strategy. In the all-on strategy, the majority of base stations operate at the minimum power with minor differences in the distribution of higher power consumption levels, with respect to the switch-on/off scheme.

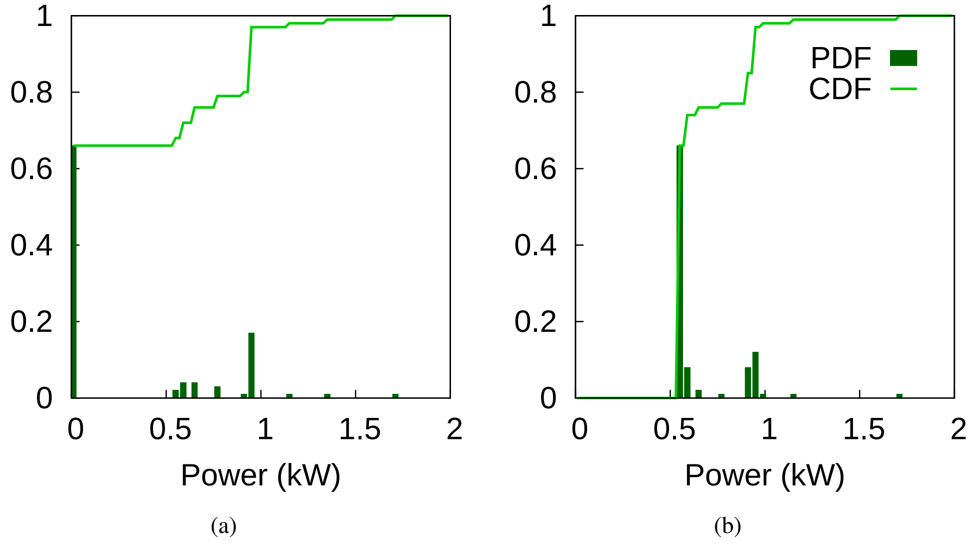


FIGURE 6.4: Distributions of BSs power levels in the geographical coverage configuration, with (a) the switch-on/off scheme and (b) the all-on strategy.

6.5 Demand accommodation with geographical coverage configuration

Once we derive the BSs power configuration, that allows to ensure the geographical coverage over the city, we aim at checking the time intervals for which the configuration allows to satisfy users demand. We do so by trying to solve the following problem, in which we optimize the allocation of resources, over GSM frames. For a certain frame, if the problem is feasible and we get a solution, then the geographical coverage configuration is enough to accommodate the demand. Otherwise, it is not: additional BSs need to be switched on in order to satisfy users demand.

In our problem, the set of y_{ij} is a set of integer decision variables, each representing the number of radio resources that shall be allocated by BS j to serve users in area i. In this problem, we define $\mathcal{A}_j$ as the set of areas associated to BS j and $\mathcal{B}_i$ as the set of BSs that can cover area i. These two sets are taken as an input from the previous problem.

$$\min \sum_{j \in \mathcal{B}} \sum_{i \in \mathcal{A}_j} y_{ij}$$

$$\sum_{i \in \mathcal{A}} y_{ij} \leq R_j \quad \forall j \in \mathcal{B} \tag{6.5}$$

$$\sum_{j \in \mathcal{B}_i} y_{ij} \geq f_i \quad \forall i \in \mathcal{A} \tag{6.6}$$

Constraint (6.5) insures that the limit on the number of radio resources for each BS, R_j, is not exceeded. Equation (6.6) imposes the constraint on the radio resources allocated for each area, such that the demand in the area can be managed. f_i represents the total demand in area i, in terms of number of calls.

We solve our resource allocation problem, over the 10 different communes, for the six hourly frames generated, as explained in Chapter 2. We show, in Fig. 6.5, the percentage of attempts, over all regions, for which the problem is feasible. We notice that these results are correlated to the actual occupation of resources in the network, shown in Fig. 6.1. In particular, we can observe that the problem leads always to a feasible solution during the late night hours, between 1:00 and 6:00. This means that switching-on base stations that ensure full geographical coverage is sufficient to accommodate the traffic load at those times in the whole city. During the day, we observe a more variable behavior. As more resources are occupied, it is less likely that the geographical configuration allows to accommodate the demand of users. However, the problem has absolutely no solution for highly occupied hours, ranging from 9:00 until 16:00.

In Fig. 6.6, we take a closer look at our results and separate them, with respect to the different communes. Interestingly, there are two communes where the geographical configuration can hold for longer hours, than others, with high percentages of feasible cases. These are the

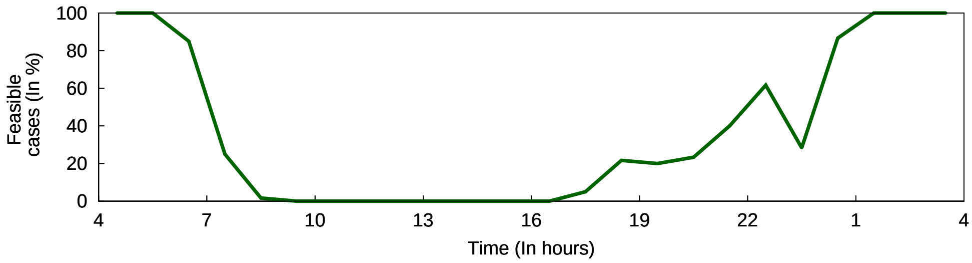


FIGURE 6.5: Percentage of feasible cases, i.e. feasible solving trials for the resource allocation problem over the day. A feasible case represents a successful solving attempt over one commune, for one out of six hourly frames, by considering different BS capacities.

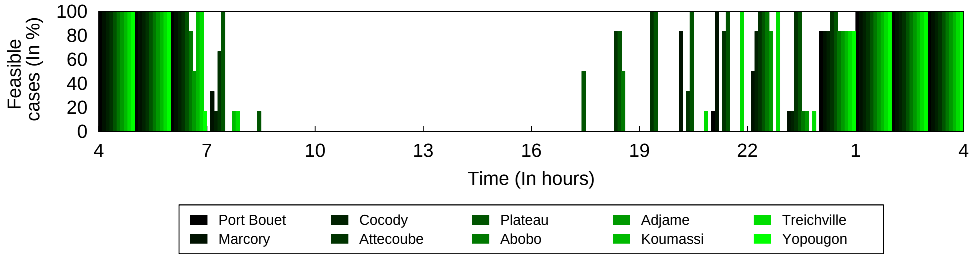


FIGURE 6.6: Percentage of feasible cases, i.e. feasible solving trials for the resource allocation problem over the day, obtained separately for each commune. A feasible case represents a successful solving attempt, for one out of six hourly frames, by considering different BS capacities.

communes of Plateau and Attecoube, with antennas located in central positions in the city, representing working areas, where human activity significantly drops, earlier than in other regions, as people leave work and go back home.

6.6 Load-adaptive BS power control

6.6.1 Load-adaptive BS power scheme

While the results presented in the previous section allow to understand how long the geographical configuration can hold users demands, they do not allow to explore the possibility of saving power at other times. Thus, we attempt to explore different options that would let us do so. A first option that we consider is to combine the two previous problems in one that can provide the minimum power consuming configuration needed to provide a full geographical coverage over the network while accommodating the demand of users at each instant.

One can think that this complete model could replace the previous two. However, this is not the case from the perspective of optimization solvers. In fact, for one traffic condition, there can be multiple equivalent optimal solutions, out of which the solver would randomly pick one. Accordingly, even if the optimal solution, derived over a certain frame, would hold for the following one, there is no guarantee that the solver would choose it.

Even more, as the solver will treat each solving attempt independently from the one over the previous frame, it will not take into account the stability of configurations. As an example, even if a minor modification to the configuration at a certain frame allows to fulfill the demand over the following frame, the solver will not necessarily consider it, and might derive a completely different solution.

Still, we formulated the complete problem and attempted to solve it, as it is capable of providing the exact power consumption needed. Unfortunately, we were not able to obtain a solution in a reasonable amount of time, due to the high complexity of the problem, even when operating over a set of tens of BSs. We thus resorted to a heuristic method, that allows to tune the BS power consumption to users demands at all times of the day.

The proposed strategy operates over the complete set of base stations $\mathcal{B}$, and thus does not consider BSs separately over different communes, as we had to, in the previous sections. However, it relies on the geographical BS configuration determined in Sec. 6.4 and considers that each BS j, switched-on according to the solution, will remain on at all times and will consume at least the minimum power suggested. We refer to this set of always switched-on BSs as $\mathcal{B}_{geo}^{on}$. Moreover, conversely to the previous section, we assume that, at each moment, a user chooses to get associated to the closest BS that can cover him. We attribute to each BS j a subset of BSs $\mathcal{N}_j \subseteq \mathcal{B}$, defined as its neighbors, according to the Voronoi diagram. BSs in $\mathcal{N}_j$ follow an ascending order based on their separation distance from BS j. We define a threshold α, representing a percentage of demand, allowing us to assess when a BS is close to becoming overloaded or underutilized. We use it in order to decide when to trigger a power control decision, as described next.

Algorithm 1 summarizes the operations of the proposed method, detailed next. At each time instant, we evaluate, for each switched-on BS j, its total demand

$$d_j = \sum_{i \in \mathcal{A}_j^{vor}} f_i, \tag{6.7}$$

where $\mathcal{A}_j^{vor}$ is the set of areas i, belonging to the Voronoi cell of BS j. This operation maps to lines spanning between Line 2 and Line 7, in Algorithm 1.

Depending on the demand at BS j, we can either decide to: *i*) trigger a power increase decision to one of its neighbors, as outlined from Line 8 to Line 17, in Algorithm 1; *ii*) trigger a switch-off decision for BS j, as indicated in the algorithm from Line 18 to Line 22; or *iii*) do nothing.

Neighbor power increase. We trigger a power increase decision for a neighboring BS k to BS j, if $d_j > (1 - \alpha)R_j$. The power increase is performed for the closest neighboring BS, whose power level is not at its maximum, $P_k^{c,max}$, through the function IncreasePower(k). The main idea behind this choice is that the closest BS k, to BS j, would be the one allowing to offload the largest amount possible of traffic from BS j. After increasing the power of a BS k, we update the Voronoi diagram over the network through function UpdateVoronoi(), which also updates the demand over all BSs. We keep on repeating this process over all neighboring BSs in $\mathcal{B}_j$, until we are able to offload the demand of BS j, to a value lower than $(1 - \alpha)R_j$.

Switch-off decision. Instead, if the demand of BS j goes below αR_j, we consider it to be underutilized. Thus, we switch it off, as long as its neighboring BSs can accommodate the resulting increase in load. The function SwitchOff(j) verifies whether this is possible or not and returns the value of P_j^c. The latter becomes equal to zero if the switch-off decision can take place. Otherwise, it maintains its initial value. Then, if BS j is switched-off, we update the Voronoi diagram, update the demand over all BSs and offload the traffic of BS j to its neighbors. We recall that we do not allow to switch off any BS j from the set of geographical coverage base stations.

6.6.2 Load-adaptive BS power configurations

We test this strategy for various values of α, over the typical working day users demand. For each case, we report in Tab. 6.3 the corresponding power gains and power state changes, representing

```
for $j \in \mathcal{B}$ do
    if $P^c_j > 0$ then
        $d_j = \sum_{i \in \mathcal{A}^{vor}_j} f_i$;
        while $(d_j > (1-\alpha)R_j)$ do
            for $k \in \mathcal{N}_j$ do
                if $P^c_k < P^{c,max}_k$ then
                    IncreasePower($k$);
                    UpdateVoronoi();
                    Break;
                end
            end
        end
        if $(d_j < \alpha R_j$ and $j \notin \mathcal{B}^{on}_{geo})$ then
            $P^c_j$ = SwitchOff($j$);
            if $P^c_j = 0$ then
                UpdateVoronoi();
            end
        end
    end
end
```

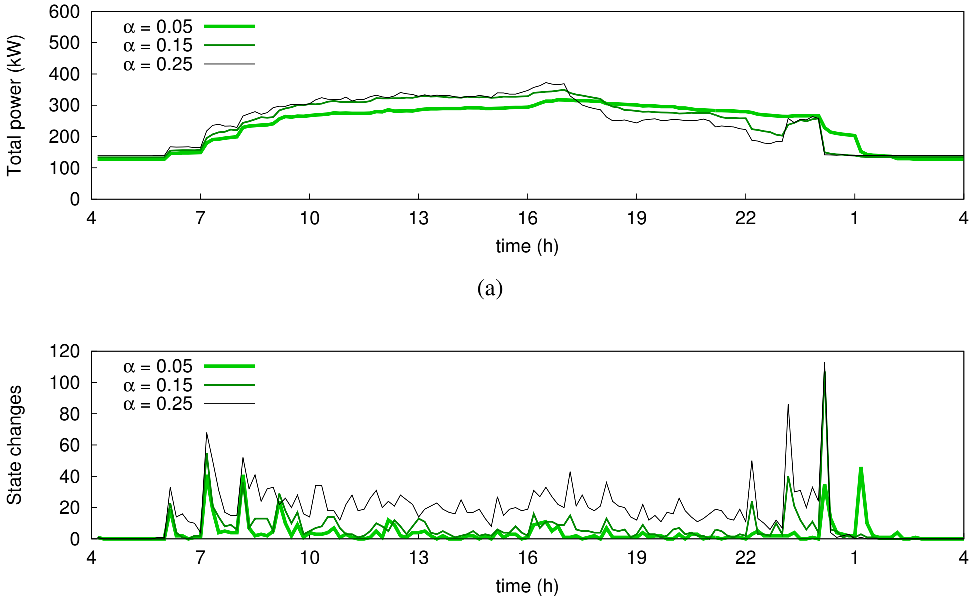
Algorithm 1: Load adaptive BS power control algorithm.

	Power gain	State changes
$\alpha = 0.05$	26 %	535
$\alpha = 0.1$	27.3%	981
$\alpha = 0.15$	28.1%	2540

Table 6.3: Consumed power gain and power state changes, obtained over the whole day, with the load-adaptive power assignment scheme, for various values of the power control threshold α.

power increase and power decrease operations over the whole network throughout the day. The power gain is accounted for by considering the total power consumption obtained based on our scheme, with respect to the case when all switched-off BSs, at each time instant, are on and consume a minimum amount of power. We observe that, for the different values of α, we obtain similar gains of almost 27%. However, we notice that α has an important impact on the global stability of the system: a higher value of α can lead to four times more power state changes in the system, translating into more system reconfigurations.

Fig. 6.7(a) reports the total consumed power, with respect to time, for each individual GSM frame. Overall, power levels remain similar for the different values of α. In particular, at low traffic hours, i.e. between 1:00 and 6:00, and for all values of α, the total consumed power converges towards the same level, corresponding to the total power consumed by the geographical coverage configuration, sufficient to satisfy users demand at those times. As the traffic increases and reaches very high values during the day, a smaller α implies less consumed power. That



(a)

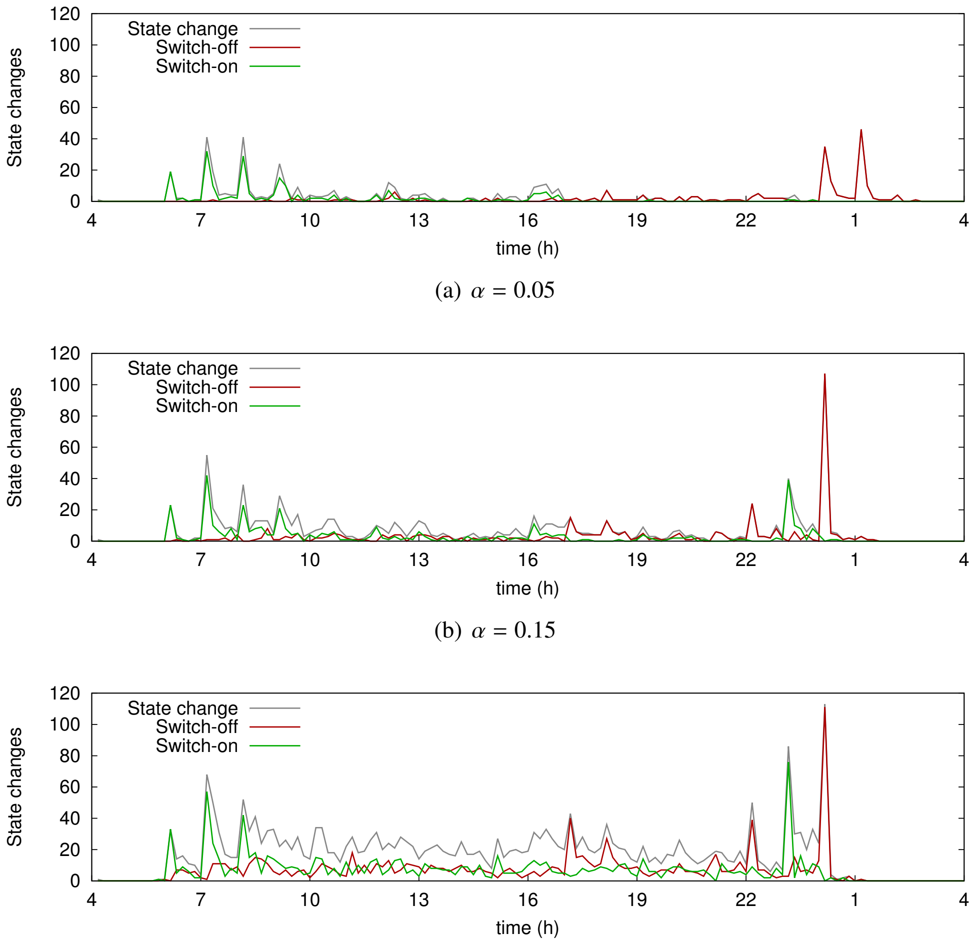
(b)

FIGURE 6.7: Consumed power (a) and power state changes (b), with respect to time, obtained with the load-adaptive power assignment scheme, for different values of α.

is due to the fact that, higher per-BS demands are required to trigger a power increase decision. Later in the day, as the traffic starts to decrease, the opposite behavior is detected, i.e. a smaller α leads to higher power consumption. The reason is that a higher α more easily triggers a switch-off decision and thus leads to smaller power consumption.

Fig. 6.7(b) complements these observations by unveiling the occurrence of power state changes over the day, for the various values of α. Clearly, we notice that a higher α leads to more reconfigurations at all times of the day, except hours with very low demand. Additionally, we notice, that for all values of α, most of the system changes occur as traffic switches between low and high levels, i.e. around 7:00 and midnight.

We take a closer look at these state changes by separating switch-on and switch-off decisions, as they relate to the most significant energy consuming decisions. Fig. 6.8 plots the corresponding results for the diverse values of α, together with the total state changing decisions. These plots confirm again the fact that a smaller value of α leads to a more stable system, with clear system reconfigurations that follow the traffic pattern suggested in Fig. 6.1. In particular, if we focus on the case of $\alpha = 0.05$, we start from the geographical coverage configuration at 4:00, which remains sufficient to accommodate users demands until 6:00, then the system undergoes significant portion of switch-on decisions that persist until a high stable traffic is reached at around 10:00. After that, the system remains relatively stable, with a few switch-on decisions and minor switch-off decisions until 17:00. Then, as the traffic starts to decrease, switch-on decisions



(a) $\alpha = 0.05$

(b) $\alpha = 0.15$

(c) $\alpha = 0.25$

FIGURE 6.8: Detailed power state changes with respect to time, for the load-adaptive power assignment scheme, for different values of α.

are not triggered anymore, instead switch-off decisions emerge, with notable peaks around midnight, marking the significant shift in users activity, driving the system back to the geographical coverage configuration, as of 1:00.

Higher values of α trigger very similar behaviors at arterial traffic switching hours, i.e. around 7:00 and midnight, however, they yield additional state changes which are not necessary in the system. This can be especially observed for $\alpha = 0.25$, with oscillating switch-on and switch-off decisions, even during stable high traffic hours.

These results underline the impact that α has over the system. While it does not have a significant impact on the power savings, it can significantly affect the stability of the system. Our evaluation point towards $\alpha = 0.05$, with the most stable system. We plot in Fig. 6.9 the corresponding total consumed power, as well as the total consumed power when all switched-off BSs are on at minimal consumption power. The figure highlights the fact that most of the gains are obtained for low traffic hours.

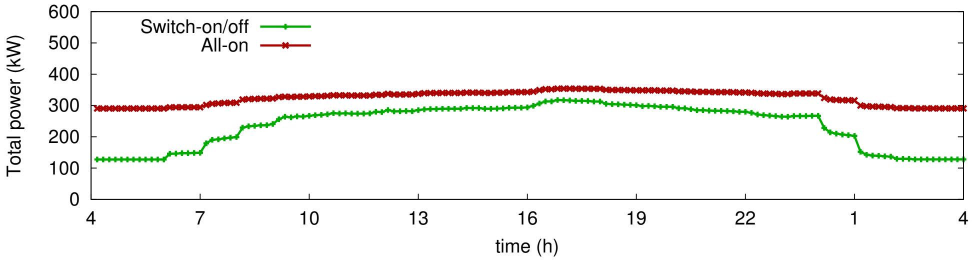


FIGURE 6.9: Consumed power for $\alpha = 0.05$ with respect to time, for the load-adaptive power assignment scheme, compared to the case when all switched-off BSs are on.

6.7 Extension to other standards

We recall that our models are adapted to GSM systems, because at the time of the study we only acquired GSM calling information for our evaluation. However, they can easily be adapted to more advanced standards. In particular, for the optimization models, to cover the LTE standard, one can rely on Shannon's formulation and simply replace Equation (6.6) by

$$\sum_{j \in \mathcal{B}_i^{max}} y_{ij} W Log_2(1 + SINR_{ij}) \geq f_i \quad \forall i \in \mathcal{A}. \tag{6.8}$$

In this case, y_{ij} would represent the number of LTE time-frequency blocks attributed from BS j to users in area i. W is the bandwidth for one LTE radio resource time-frequency block, which can be considered constant for simplicity. f_i represents the total throughput demand in area i. $SINR_{ij}$ is the signal-to-interference-plus-noise ratio perceived by a user in area i, with respect to the signal received from BS j. The left part of constraint (6.8) provides the sum of theoretical throughputs that can be attained, based on Shannon's formula in area i.

Based on similar assumptions, the load-adaptive strategy can also be extended to LTE systems.

6.8 Conclusion

In this chapter, we employed mobile traffic datasets to evaluate achievable power savings in a cellular network, resulting from load-adaptive BS power control schemes, that account for geographical coverage and demand accommodation constraints. We adopt a methodology that combines optimization and heuristic techniques. We first introduce an optimized power consumption BS configuration, that ensures geographical coverage over the network. Then, we assess its capability to satisfy users demand, over a whole day. Finally, we propose a methodology that allows to adapt this configuration to all traffic states, so as to assess possible savings at all times of the day. Our results indicate that important savings, of almost 27%, can be achieved over real-world traffic patterns.

Chapter 7

Dynamic Cloud Radio Access Networks and User Mobility

7.1 Introduction

In the previous chapter, we exploit mobile traffic datasets to tackle the problem of reducing power consumption, over the radio access part of current cellular networks. As discussed, this problem is of major importance particularly for our current highly dense networks. Nevertheless, another important challenge also results from such deployments, when it comes to mobile customers. In fact, for static users, dense deployments are beneficial on average, from a capacity point of view. However, the situation is problematic for mobile users, whose movements translate into an increase of handovers over the network, with respect to a sparser deployment.

From a network perspective, this reflects additional signalling traffic, at a time when operators are hardly able to handle the explosion of users traffic. From the viewpoint of mobile users, this is perceived as a degradation in the quality of service, due to the connection disruption, coming along the hard handovers procedures, widely implemented in cellular systems.

Within this context, considering a flexible network topology, that can adapt to the mobility of users and their traffic demand can be very beneficial. In particular, a network architecture that would allow to perform baseband processing of several cells over one equipment, can result in less handovers over the network, especially if designed so as to take into account human mobility patterns. The newly envisioned Cloud-Radio Access Network (Cloud-RAN) [114, 115] topology provides this possibility, thanks to its promising highly adaptive capabilities.

In this chapter, we shed light on the potential of the dynamic Cloud-RAN topology for handling the mobility of users. We start by introducing the general Cloud-RAN architecture, in Sec. 7.2. Then, we review, in Sec. 7.3 previous works, related to our study. Sec. 7.4 and Sec. 7.5 present

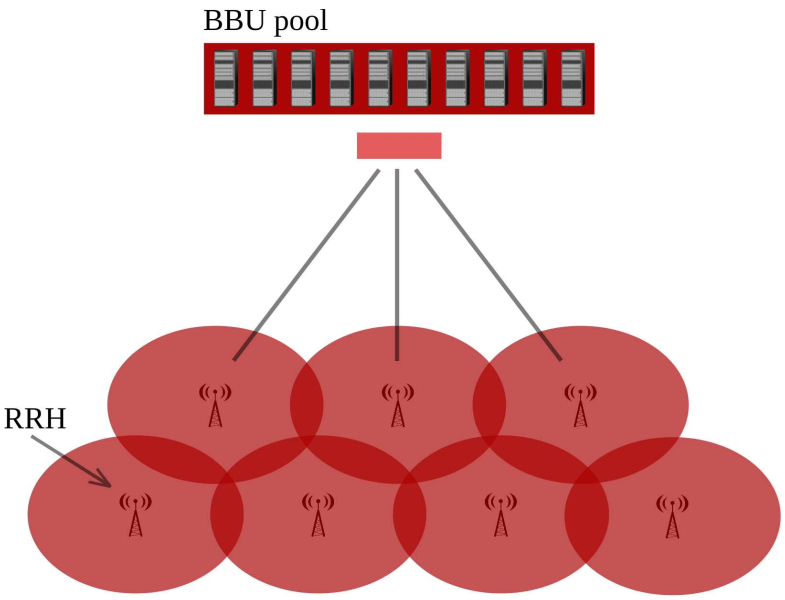


FIGURE 7.1: Cloud-RAN architecture.

the methodology that we propose to adapt the Cloud-RAN topology to the dynamics of users. Finally, in Sec. 7.7, we conclude the chapter. We remark that we summarize in Tab. 7.1 all variables employed throughout the chapter.

7.2 Architecture of Cloud-RAN

In this section, we briefly cover the evolution of base station (BS) systems, progressively paving the way towards the Cloud-RAN architecture, followed by a short overview of general aspects relating to the concept of Cloud-RAN.

7.2.1 Evolution of BS system

In order to allow users to communicate over the network, BSs perform some baseband processing, including, among others, channel coding and signal modulation, as well as some radio functionalities, such as frequency filtering and power amplification. Traditionally, these radio and baseband processing operations have been implemented in the same BS equipment, with the radio module connected through coaxial cable to the baseband processing module. That was the case for 1G and 2G BSs.

Later, in addition to initial BS systems, more flexible ones have been introduced, within the 3G architecture, with a clear split of radio and baseband processing operations over geographically separated entities, Remote Radio Head (RRH) and Baseband Unit (BBU), connected through fiber. This has allowed operators to explore the benefits of local BBU centralization, reducing expenses, in terms of site rental and maintenance costs.

Nevertheless, associations between BBU and RRH equipments remained static, with one BBU handling the demand of only one RRH, that it serves at all times. Thus, modifying the BBU-RRH mapping, over time, could not be achieved with such systems, while granting a higher degree of flexibility in the overall system.

The Centralized-Radio Access Network (Centralized-RAN) architecture, an early version of the Cloud-RAN concept, makes this dynamic mapping possible, with its centralized processing design. The Centralized-RAN includes a set of geographically distributed RRHs, connected through fiber or microwave connections to data centers. The latter will include BBU pools, that would facilitate cooperations and sharing of resources among different BBU equipments. More important is the fact that with such an architecture, a BBU would be able to cover several RRHs, with associations that can change over time.

This concept is even pushed further with the architecture of Cloud-RAN, providing simpler ways to manage the network, by employing network virtualization techniques. We illustrate the Cloud-RAN architecture in Fig. 7.1.

7.2.2 Potential and challenges of Cloud-RAN

The Cloud-RAN architecture, holds a lot of potential for future 5G networks. First, as shown by previous analyses in Chapter 3, and our own in Chapter 4, mobile traffic is not uniform over space and time and undergoes significant variations over both dimensions. Typically, networks are dimensioned to accommodate peak traffic. This obviously results in wasting processing resources, most of the time. Due to its resource sharing capability, a well-dimensioned Cloud-RAN architecture would enable a better utilization of resources, that can adapt to traffic variations.

Second, with its centralized architecture, Cloud-RAN requires less cooling resources. This translates into important savings in energy consumption and thus lower electricity expenses for operators.

Third, Cloud-RAN also enables simpler operations for managing the network. Examples include multistandard operations, cell cooperations, upgrading the network, and expanding it. This is especially true for the case of Cloud-RAN, where all these functionalities can be completed through simplified software functions. For instance, typical costly distributed hardware upgrades can be performed by simple software updates in the BBU pool.

While Cloud-RAN holds promising potential for cellular networks, from several perspectives, a number of challenges remain to be addressed. The centralization of BBU processing implies a large overhead over the fronthaul links between RRHs and data centers. There is thus a need for designing a cost-efficient high-bandwidth transport network between them.

At the level of data centers, Cloud-RAN requires virtualization techniques that would enable the functionalities of BBUs, over virtual machines. These techniques need to operate in real-time and meet the delay constraints imposed by cellular network standards.

Finally, dynamic network management solutions need to be devised, so as to enable optimal network operations. In our work, we contribute to ongoing efforts in this direction, and focus particularly on the dynamic management of the Cloud-RAN topology. We review corresponding related works, in the next section.

7.3 Literature review

Several previous works have studied dynamic Cloud-RAN topology reconfiguration solutions, with time-varying BBU-RRH associations. Some of these studies analyzed the potential of such techniques.

Namba *et al.* [116] evaluate possible equipment savings, in terms of number of BBUs, in a Cloud-RAN architecture, with respect to a typical RAN. To do so, they calculate how many BBUs are needed in the network in order to accommodate traffic load in two simulated traffic scenarios: day and night. Their results indicate that, with respect to a typical RAN, the number of BBUs can be reduced by a portion of 75 to 81%.

Similarly, Checko *et al.* [117] calculate the number of BBUs required at each instant of the day to accommodate the demand of users. In their work, they simulate mixed traffic profiles in residential, work and commercial areas. Their results confirm the conclusion by Namba *et al.* [116], as they observe that a Cloud-RAN architecture requires four times less BBUs than current distributed configurations.

Later, Checko *et al.* [118] extend the scope of these studies by analyzing how BBU equipment savings would change with respect to diverse mixtures of traffic profiles. More precisely, they calculate the number of BBUs for varying ratios of residential and working traffic patterns. They conclude that the highest gains are achieved when centralizing BBUs with 20-30% of cells covering office areas, with the rest covering residential zones.

Liu *et al.* [119] stress the benefits of a Centralized-RAN architecture based on real-world experiments conducted over a pool of 4 BBUs and 4 antennas. In addition to savings in BBU equipments, they consider gains in the network performance in terms of throughput. More precisely, they evaluate achievable throughput with different BBU-RRH static mappings: one-to-one, one-to-many and hybrid scheme, for both mobile and static users. Their results indicate that a reconfigurable system holds the capability of balancing between performance gains and infrastructure costs.

Other works explored benefits of dynamic schemes, by focusing on computational gains. Madhavan *et al.* [120] evaluate computational gains, defined as gains in CPU cycles, resulting from performing medium access layer functionalities of a WiMAX network in a centralized cloud. They observe that a significant gain, scaling linearly to the network size, can be obtained. Moreover, they notice the gain increases as the traffic intensity becomes higher.

Werthmann *et al.* [121] evaluate computational gains, expressed in terms of number of operations per second, resulting from centralizing all base band processing. By simulating an LTE traffic scenario, based on mathematical distributions, they observe that, by aggregating 57 sectors, savings can reach 25%. Additionally, they highlight the fact that the spatial distribution of users has an important impact on the utilization of compute resources.

In turn, Bhaumik *et al.* [122] analyze 3G mobile traffic traces over 21 cell sites and compare computational efforts, in terms of processing load resulting from coding and modulation operations, in a centralized architecture, to those in a decentralized one. They observe that a centralized architecture can imply savings of at least 22%. Furthermore, they introduce a framework to manage BBUs and schedule their operations in data centers, granting compute efforts savings of up to 19%.

Liu *et al.* [123] underline the additional benefits of the Cloud-RAN architecture, from a mobility point of view. By analyzing handover procedures in different cellular networks generations, they observe the following: with respect to GSM, Cloud-RAN grants lower connection interruption duration, with respect to UMTS it reduces signaling overheads, and with respect to LTE, it decreases handover delay as well as handover failure rate.

Previously discussed works unveil the potential of a Cloud-RAN architecture with dynamic system configurations, however, they do not propose methodologies that would allow to do that.

Namba *et al.* [124] deal with this aspect and introduce two BBU-RRH switching schemes. The first one is a semi-static algorithm, which aims at accommodating peak traffic over a whole day. The algorithm assigns RRHs to BBUs, such that the capacity limit of a BBU is not exceeded, while favorizing the assignment of neighboring RRHs over the same BBU. The second strategy provides associations that vary over time with respect to the load and resource usages. It aims at balancing usages among BBUs, while also favorizing the assignment of neighboring RRHs to the same BBU. Both strategies are evaluated over a set of 100 cells with both working and residential traffic profiles. The results indicate that the semi-static algorithm leads to 26% of gains in terms of number of BBUs. Instead, the adaptive strategy is observed to imply higher savings that can go up to 47%.

Zhu *et al.* [125] introduce a dynamic BBU-RRH scheme which aims at load-balancing BBUs, while minimizing co-channel interference among couples of RRHs associated to different BBUs. This algorithm is complemented by a scheduling strategy that allows to further cancel remaining co-channel interference. The evaluation of this scheme, over a set of 84 RRHs, with simulated traffic, shows that it can reduce 60% of power consumption, with throughput gains of more than 45%.

Wang *et al.* [126] consider another aspect, relating to the system configuration. They propose an algorithm to allocate frequency resources, among different cell zones, in a Cloud-RAN architecture. The algorithm relies on a graph coloring methodology [127] and aims at allocating resources by taking into account the traffic demands, while avoiding inter-cell interference. By evaluating the scheme in a simulative environment with several cells, the authors observe that it is capable of reducing the energy consumption in the network and improving the spectrum utilization.

Overall, these works have stressed the potential of a dynamic Cloud-RAN topology and introduced dynamic BBU-RRH association algorithms, that can allow to do that. These studies have been conducted mostly in simulative environments [116–118, 120, 121, 124–126], over a relatively small scale and by considering a limited set of traffic profiles, assumed to be representative of e.g. residential, working, and commercial areas. While traffic of antennas, located in areas with similar land-use characteristics present similar characteristics, they can still present notable variations in traffic intensity, as traffic is not completely homogeneous over space.

Moreover, these works have focused on the benefits of associating RRHs with dissimilar traffic patterns to the same BBU. In reality, operators would prefer to favorize associations of closely located RRHs over the same BBU, as this can be useful for cooperative management schemes. However, one would expect RRHs that are geographically close to present similar traffic patterns.

Real-world traffic traces have been employed in [119, 122]. Nevertheless, they remain limited to only a small number of RRHs.

Besides, these works mainly explore the variability in demand to underline achievable gains in the network. Not much attention has been paid to mobility aspects, although of notable importance. The work by Liu *et al.* [123] remains an exception there. Still, the authors only conduct a qualitative study, neglecting major cellular network constraints such as capacity.

As such, in this work, we propose a dynamic BBU-RRH switching scheme which aims at adapting associations to users mobility, while still satisfying their traffic demand. Conversely to previous works, we evaluate our scheme using real-world mobile traffic profiles, over a large scale.

7.4 System modeling

Our objective, in this work, is to find dynamic BBU-RRH associations, that allow to minimize handovers in the network, while still satisfying the demand of users. To model this problem, we adopt a graph representation of the network, allowing us to explore a graph clustering technique in order to derive a solution.

7.4.1 Static graph representation

A first possibility to model the system could be based on a series of separate graphs holding necessary information to decide on system reconfiguration. We illustrate such an option in Fig. 7.2. There, we consider snapshots of traffic, that include information about users demands over a significant time interval Δt, representing the smallest scheduling time interval. Δt can be defined with respect to the cellular network standard. As an example, for the case of GSM, it can be considered as a GSM frame. We employ $\mathcal{T}$ to refer to a set of traffic snapshots t over a certain observation period. Each couple of consecutive snapshots that we consider in $\mathcal{T}$ are separated by a time interval $\Delta t'$, with $\Delta t \leq \Delta t'$. We use i^t to represent an RRH appearing in the network over snapshot t and refer to the complete set of RRHs as $\mathcal{R}^t$ over snapshot t. We denote the demand of i^t over snapshot t as d_i^t. We consider that a couple of RRHs i^t and j^t are neighbors if the corresponding cells in the Voronoi diagram of the cellular network share a common border. We construct a graph $G_t(\mathcal{R}^t, \mathcal{E}^t)$, over the set of RRHs $\mathcal{R}^t$. The set $\mathcal{E}^t$ is a set of edges e_{ij}^t, linking neighboring RRHs i^t and j^t. We consider that a couple of consecutive snapshots are separated by a time interval $\Delta t'$, referring to the smallest time interval for system reconfigurations. Each edge e_{ij}^t is assigned a weight h_{ij}^t representing the total number of handovers between RRH i and RRH j over the time interval $\Delta t'$ spanning between snapshot $t-1$ and snapshot t.

Our objective is to find a set of clusters $\mathbb{C}^t$, where a cluster $\mathbb{C}_k^t \in \mathbb{C}^t$ represents a set of RRHs associated to one BBU k. The optimal clustering solution allows to solve the following optimization problem:

$$\min \sum_{i^t \in \mathbb{C}_k^t, j^t \in \mathbb{C}_l^t} h_{ij}^t$$

$$\sum_{i^t \in \mathbb{C}_k^t} d_i^t \leq cap_k \quad \forall k, \forall t, \tag{7.1}$$

Variable	Significance
cap_k	Capacity limit of a BBU k over Δt
d_i^t	Traffic demand over i^t
$e_{ii}^{t,t+1}$	Edge linking RRHs i^t and i^{t+1}
e_{ij}^t	Edge linking neighboring RRHs i^t and j^t
$h_{ii}^{t,t+1}$	Handovers between i^t and i^{t+1} due to reconfiguration
h_{ij}^t	Handovers between neighboring RRHs i^t and j^t between $t-1$ and t
i^t, j^t	RRH at snapshot t
k, l	BBU equipments
l_{uv}	Edge linking node u and node v
r, s	Nodes in $\mathcal{N}^{temp}$
s_{max}	Selected neighboring node to perform a power increase in the modified Louvain algorithm
t	Traffic snapshot
u, v	Nodes in $\mathcal{N}$
w_u	Sum of weights of edges attached to node u
w_{uv}	Weight of edge l_{uv}
Δt	Duration of a snapshot
$\Delta t'$	Time interval elapsed between a couple of snapshots
ΔQ_{rs}	Modification in modularity resulting from letting node r join the community of node s
ΔQ_{max}	Maximum increase in modularity for a node with respect to its neighbors
G_t	Static graph over snapshot t
G_{temp}	Temporary graph structure employed in the modified Louvain algorithm
G	Time-varying graph over the whole observation period
H	A generic graph
Q	Modularity of a graph partitioning
W	Total sum of weights of edges in the network
$\mathcal{E}$	Set of edges in the time-varying graph
$\mathcal{E}^t$	Set of edges e_{ij}^t over snapshot t
$\mathcal{E}^{t,t+1}$	Set of edges $e_{ii}^{t,t+1}$ in G
$\mathcal{L}$	Set of edges in graph G
$\mathcal{L}^{temp}$	Set of edges in G_{temp}
$\mathcal{N}$	Set of nodes in graph G
$\mathcal{N}^{temp}$	Set of nodes in G_{temp}
$\mathcal{N}_r^{temp}$	Set of neighboring nodes to node r in G_{temp}
$\mathcal{R}$	Set of RRHs over the whole observation period
$\mathcal{R}^t$	Set of RRHs over snapshots t
$\mathcal{T}$	Set of traffic snapshots
$\mathbb{C}_k^t$	Cluster of RRHs associated to BBU k
$\mathbb{P}_m$, $\mathbb{P}_n$	Partition over graph H
$\mathbb{C}$	Set of clusters over the time-varying graph
$\mathbb{C}^t$	Set of static clusters $\mathbb{C}_k^t$ at snapshot t
$\mathbb{P}$	Partitioning over graph H

TABLE 7.1: Variables employed throughout the chapter.

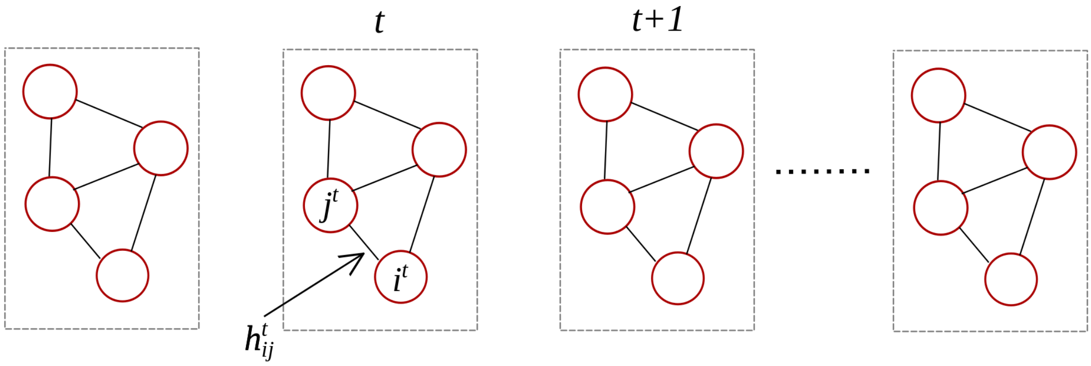


FIGURE 7.2: Static graph representation.

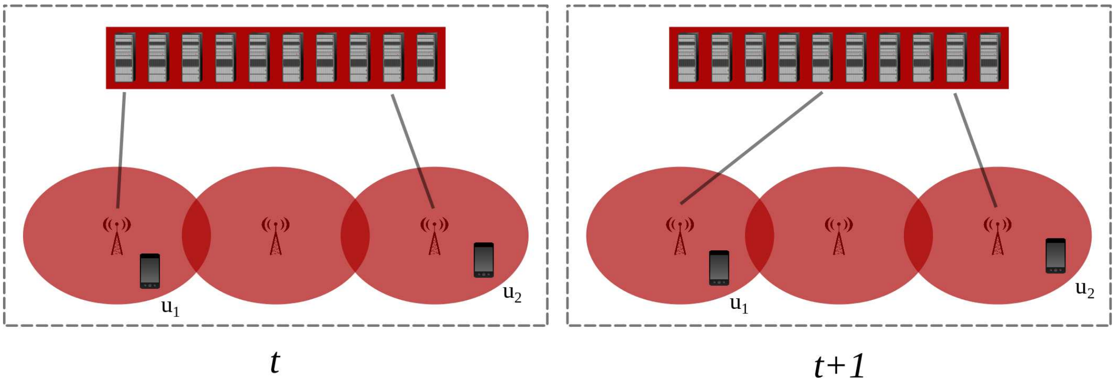


FIGURE 7.3: Illustration of additional handovers resulting from differences in system configurations at successive snapshots t and $t + 1$. Static user u_1 encounters a handover, while static user u_2 does not.

where cap_k represents the capacity limit of a BBU k, over Δt. While solving this problem allows to derive a solution adapted separately for different network behaviors, it fails to capture the continuity of traffic. This can result in frequent system reconfigurations which may not be beneficial for the global system performance. In fact, each system reconfiguration with modified RRH-BBU associations introduces handovers in the network even for static users. Fig. 7.3 shows an example that illustrates such a case, for one user. Suppose that at snapshots t and $t + 1$, we obtain the BBU-RRH associations portrayed in the figure. Focusing on the static user u_1, who remains connected over both snapshots to the same RRH, he will encounter a handover as the system is reconfigured. The reason behind this is that the corresponding RRH switches to a new BBU. Instead, this will not be the case for user u_2, whose corresponding RRH remains associated to the same BBU over both snapshots. This behavior is much more critical Clearly, such operations can have an important impact on the quality of experience perceived by users, an important aspect that has not been considered by previous works.

In our study, we resort to a dynamic representation described in the following subsection.

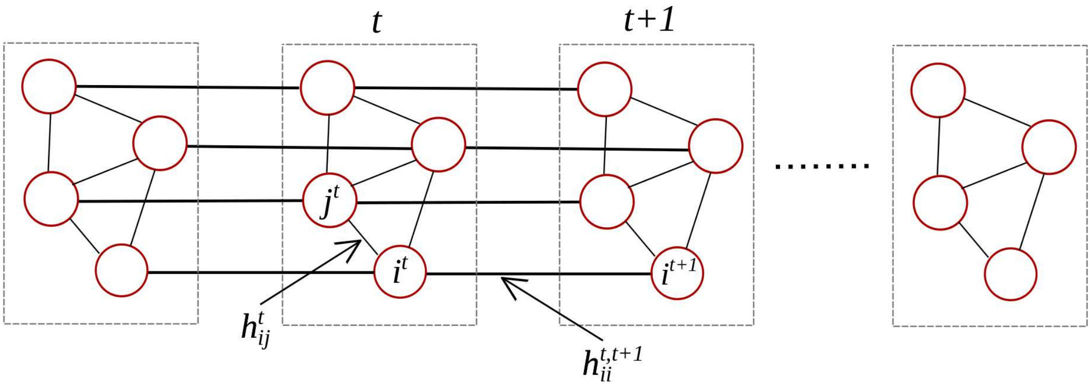


FIGURE 7.4: Time-varying graph representation.

7.4.2 Time-varying graph representation

We employ a time-varying graph model $G(\mathcal{R}, \mathcal{E})$ to capture the system dynamics, as shown in Fig. 7.4. The graph is built over the complete set of RRHs over the whole observation period $\mathcal{R} = \bigcup_{t \in \mathcal{T}} \mathcal{R}^t$. A set of edges $\mathcal{E} = \bigcup_{t \in \mathcal{T}} \{\mathcal{E}^t \cup \mathcal{E}^{t,t+1}\}$ links RRHs in $\mathcal{R}$. $\mathcal{E}^t$ holds the same meaning as in the static graph, and $\mathcal{E}^{t,t+1}$ is a set of edges $e^{t,t+1}_{ii}$ linking node i_t to node i_{t+1}. We assign a weight to edge $h^{t,t+1}_{ii} = d^t_i$, representing the total number of potential handovers resulting from a reconfiguration that associates i^t and i^{t+1} over two different BBUs.

Based on that, our objective is to group RRHs, over time, into a set of clusters $\mathbb{C}$, such that a cluster $\mathbb{C}_k \in \mathbb{C}$ constitutes a set of RRHs which will be associated to the same BBU k, over time. The optimal clustering corresponds to the solution of the following optimization problem:

$$\min \sum_{i^t \in \mathbb{C}_k, j^t \in \mathbb{C}_l} h^t_{ij} + \sum_{i^t \in \mathbb{C}_k, j^t \in \mathbb{C}_l} h^{t,t+1}_{ii}$$

$$\sum_{i^t \in \mathbb{C}_k} d^t_i \leq cap_k \quad \forall k, \forall t \tag{7.2}$$

7.5 Dynamic Cloud-RAN associations scheme

For our problem, we employ a modified version of the Louvain method [128], which we adapt to our constraints. The Louvain method aims at extracting clusters from large graphs, with the objective of maximizing the modularity (Q). In the following, we start by presenting the Modularity over a clustering in a generic graph. Then, we describe the different steps of our modified version of the Louvain method.

7.5.1 Modularity

We consider a generic graph $H(\mathcal{N}, \mathcal{L})$, built over a set of nodes $\mathcal{N}$, and linked by a set of edges $\mathcal{L}$. A couple of nodes u and $v \in \mathcal{N}$ are linked by an edge l_{uv}, to which we assign a weight w_{uv}. Suppose that H is divided into a set of clusters or partitions $\mathbb{P}$.

The modularity for the graph partitioning $\mathbb{P}$ compares the cohesion inside partitions to the case of a random distribution of edges over partitions. It takes a value ranging between -1 and 1. A high value of modularity indicates a high cohesion of links inside each partition, with respect to the links among them. The modularity can be evaluated as follows:

$$Q = \frac{1}{2W} \sum_{u \in \mathcal{N}, v \in \mathcal{N}} \left(w_{uv} - \frac{w_u w_v}{2W} \right) \delta(\mathbb{P}_m, \mathbb{P}_n), \tag{7.3}$$

where we consider that node u belongs to partition $\mathbb{P}_m$ and node v belongs to partition $\mathbb{P}_n$.

$$w_u = \sum_{v \in \mathcal{N}} w_{uv} \quad \forall u \in \mathcal{N} \tag{7.4}$$

is the sum of weights of edges attached to a node u, and

$$W = \frac{1}{2} \sum_{u \in \mathcal{N}, v \in \mathcal{N}} w_{uv} \tag{7.5}$$

is the total sum of weights of edges, in the whole graph H. The δ-function $\delta(x, y)$ is equal to one if $x = y$ and 0 otherwise.

Optimizing the modularity is a computationally hard problem [129]. The Louvain method is a heuristic method that has been observed to outperform other known clustering methods in terms of computational efforts, while still deriving clusters with very good quality [128].

7.5.2 Modified Louvain method

We outline in Algorithm 2 the steps that we follow, in our modified version of the Louvain method, to cluster the graph. In the algorithm, we mark in red the parts that we add, representing the modifications with respect to the main Louvain method. Moreover, Fig. 7.5 highlights the major steps of the algorithm, detailed next.

The algorithm mainly operates over a graph structure $G^{temp} = (\mathcal{N}^{temp}, \mathcal{L}^{temp})$, where $\mathcal{N}^{temp}$ forms a set of nodes, and $\mathcal{L}^{temp}$ is a set of weighted edges linking them. G^{temp} is updated at each iteration.

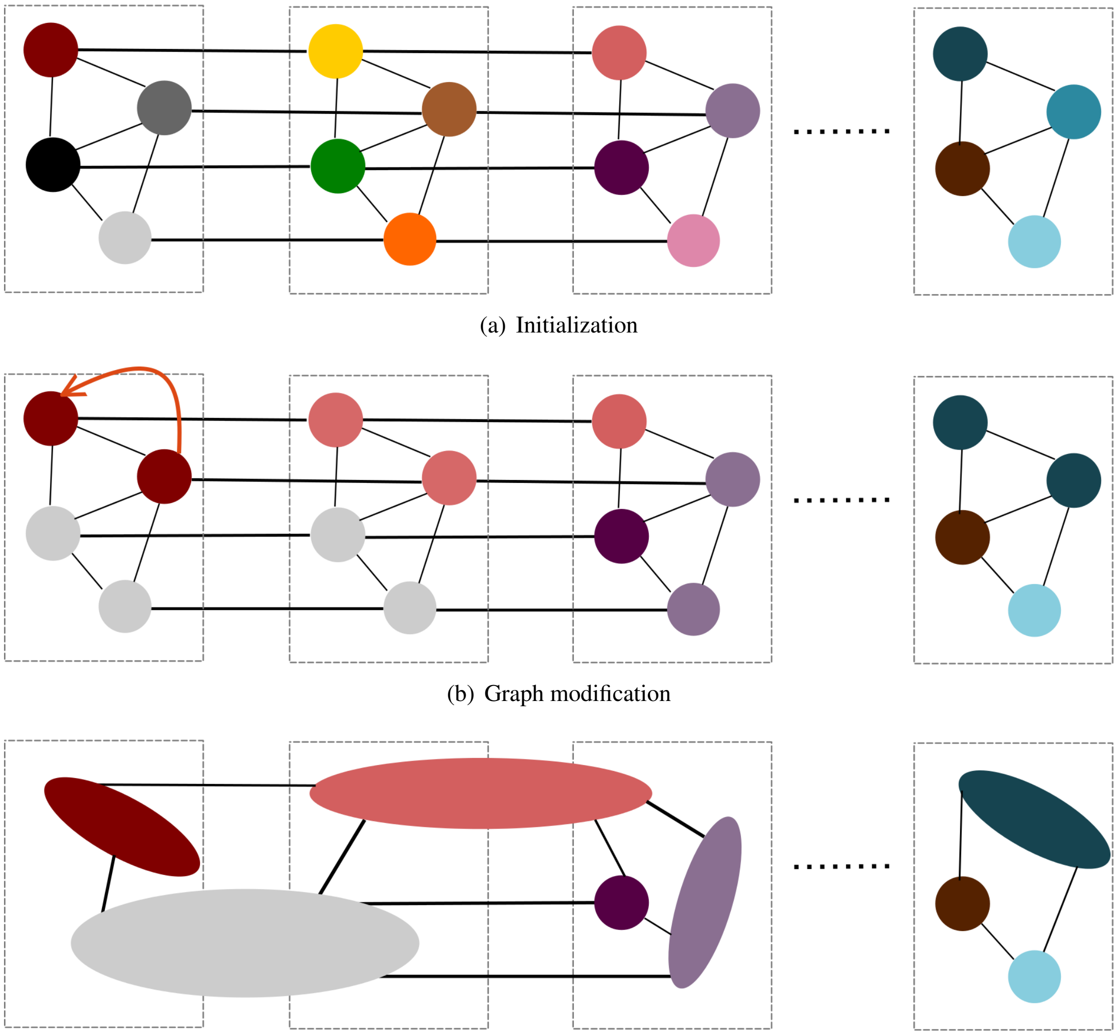
```
1  $G^{temp}, \mathbb{C}$ = Attribute($\mathcal{R}$);
2  modif = True;
3  rebuilt = True;
4  while rebuilt == True do
5      while modif == True do
6          modif = False;
7          for $r \in \mathcal{N}$ do
8              $\Delta\mathrm{Q}_{max} = 0$;
9              $s_{max} = 0$;
10             $\mathcal{N}_r^{temp}$ = ExtractNeighbors($r$);
11             for $s \in \mathcal{N}_r^{temp}$ do
12                 $\Delta\mathrm{Q}_{rs}$ = Calculate$\Delta$Q($r$,$s$);
13                 if $\Delta Q_{rs} > \Delta Q_{max}$ then
14                     verif = VerifyCapacity($r$,$s$);
15                     if verif == True then
16                         $\Delta\mathrm{Q}_{max} = \Delta\mathrm{Q}_{rs}$ ;
17                         $s_{max} = s$ ;
18                     end
19                 end
20             end
21             if $\Delta Q_{max} > 0$ then
22                 JoinCommunity($r$,$s_{max}$);
23                 modif = True;
24             end
25         end
26     end
27     rebuilt,$G^{temp}, \mathbb{C}$ = RebuildGraph($G^{temp}, \mathbb{C}$);
28     modif = True;
29 end
```

Algorithm 2: Modified Louvain method.

The algorithm starts by forming G^{temp}, by considering that initially $\mathcal{N}^{temp} = \mathcal{R}$ and $\mathcal{L}^{temp} = \mathcal{E}$. It also assigns the weight of each edge in $\mathcal{E}$, to its counterpart in $\mathcal{L}^{temp}$. Moreover, it associates to each RRH in $\mathcal{R}$ a cluster $\mathbb{C}_k$. This step is summarized in Line 1 of the algorithm.

After this initialization step, the algorithm oscillates between two steps: *i*) Graph modification and *ii*) Graph reconstruction described in the following.

Graph modification. The algorithm attempts to modify the structure of G^{temp}, as follows. Each node of the graph $r \in \mathcal{N}$ is attributed a set $\mathcal{N}_r^{temp} \subseteq \mathcal{N}$, forming its neighbors in the graph. At each step, a node r will join the community of a neighbor $s \in \mathcal{N}_r^{temp}$, which gives the maximum increase in modularity, as long as this does not violate the capacity constraint of the cluster to which belongs s. This step is repeated as long as we have positive modularity variations over the graph. Lines spanning between Line 5 to Line 26, in Algorithm 2, describe this step. There, the function ExtractNeighbors(r) returns the set $\mathcal{N}_r^{temp}$. The function CalculateΔQ(r, s), returns the modification in the modularity in case node r joins the community of node s. Function



(a) Initialization

(b) Graph modification

(c) Graph reconstruction

FIGURE 7.5: Illustration of the modified Louvain method steps over the time-varying graph. Each node takes a color, referring to the cluster to which it belongs.

VerifyCapacity(r, s) checks if letting node r join the community of node s results in overpassing the capacity limit of the corresponding BBU. If that is the case, the function returns False, and otherwise it returns True. Finally, the function JoinCommunity(r, s) lets node r join the community of s.

Graph reconstruction. Once there are no more possible modifications implying positive variations in the modularity, the algorithm rebuilds the graph, by taking communities as nodes and linking them with a new set of edges that are weighted by the sum of the total inter-cluster link weights. This step is completed through function RebuildGraph(G^{temp}, $\mathbb{C}$). The function attempts to reconstruct G^{temp}: if it obtains a graph that is the same as the one taken as input, it assigns a False value to the rebuilt variable, and returns otherwise a True value. Additionally, it provides G^{temp}, and the communities resulting from the reconstruction step.

	HO_{sav}	HO_{rem}
Low mobility	0.23	0.85
High mobility	0.24	0.46

TABLE 7.2: Percentage of handovers saved HO_{sav}, i.e. percentage of handovers removed when comparing the Cloud-RAN architecture to the traditional RAN, and percentage of handovers removed HO_{rem}, i.e. percentage of handovers eliminated from the Cloud-RAN graph.

7.6 Evaluation

In this section, we present the characteristics of our evaluation environment, followed by the results that we obtain.

7.6.1 System parameters and assumptions

We evaluate our strategy using the D4D datasets described in Sec. 2.2 with the following additional parameters and assumptions.

RRH settings. We derive our results based on snapshots of GSM per-frame users demands. This allows us to assess the potential of our methodology with respect to the virtualization of a GSM system. We refer the reader to Sec. 7.6.5 for a discussion concerning other systems. These snapshots are obtained over one day and separated by ten-minute time intervals. The per-frame demand is generated from real-world hourly per-BS traffic profiles, as detailed in Sec. 2.3.3.1, except that we skip the step of geographically distributing the demand of a BS over areas it covers. We assume that RRHs occupy the position of Orange BSs. Therefore, we consider that the demand of one RRH is that of the BS it replaces.

Handovers. We consider two handover scenarios: a low and a high mobility ones. The per-frame handovers information is obtained as described in Sec. 2.3.3.2. We aggregate this information over ten-minute time intervals, so as to assign the weights h^t_{ij} in our graphs.

BBU settings. We consider that BBUs are high capacity equipments, such that each can hold the demand of a high capacity GSM BS, with a total of 18 carrier frequencies. We assume that all BBUs are grouped in one data center, covering the whole city.

7.6.2 Handovers gain

We summarize, in Tab. 7.2, savings obtained for the two mobility scenarios over the whole day, in terms of handovers. In particular, we show the percentage of handovers saved HO_{sav}, i.e. percentage of handovers eliminated when comparing the Cloud-RAN configurations, that we derive, to the traditional RAN configurations. We notice that for the two scenarios our strategy

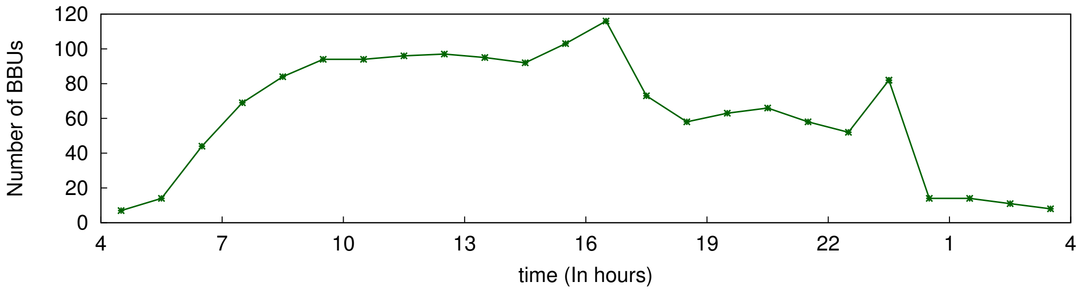


FIGURE 7.6: Average number of BBUs required over the whole day.

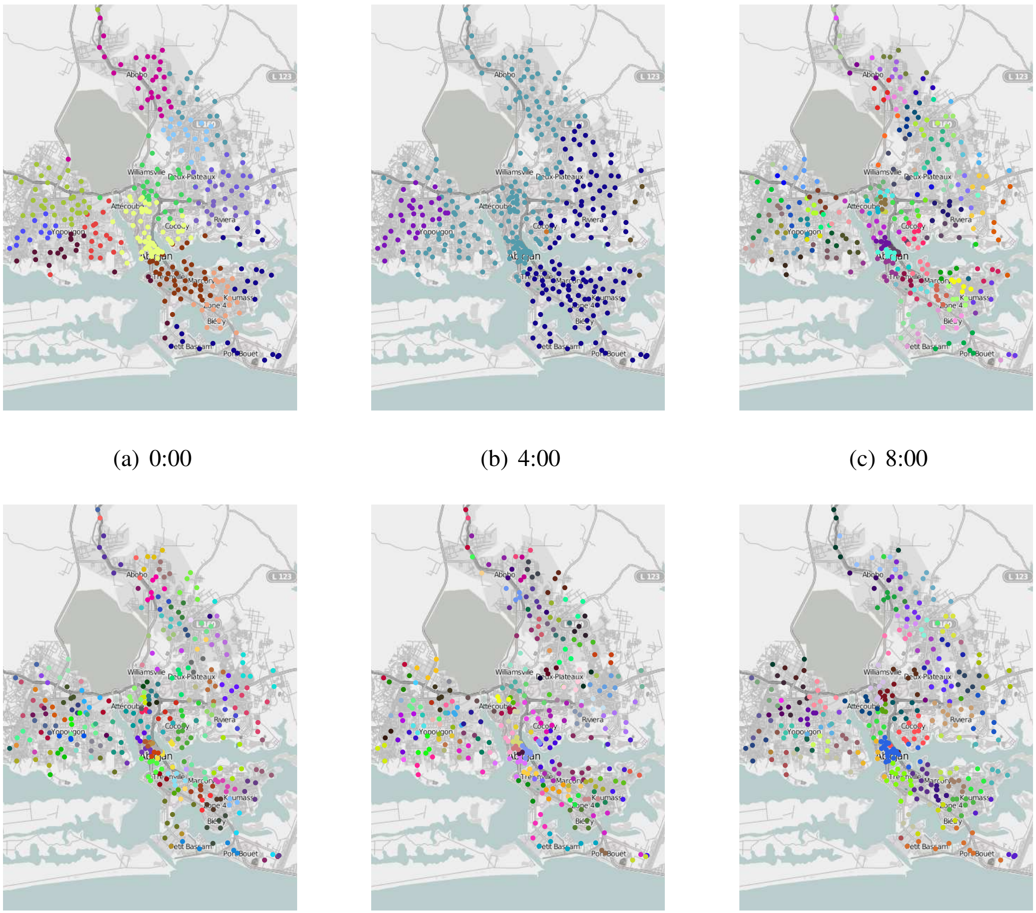
leads to very similar values of HO_{sav} of almost 24%. Tab. 7.2 also shows the percentage of handovers removed from the whole Cloud-RAN graph, HO_{rem}. We notice that for the low mobility scenario, we are able to remove 85% of the total number of handovers. This percentage is lower for the case of high mobility scenario, where we can remove 46% of all handovers. This is due to the fact that a low mobility scenario grants more flexibility when it comes to the way clusters are structured, resulting in a higher percentage of savings.

7.6.3 Global system evolution

We complement these results by checking the overall evolution of the BBUs in the network. We focus on the high mobility scenario, more challenging than the low mobility one. We plot in Fig. 7.6 the average number of BBUs, required at each hour, so as to serve the demand of users over the six hourly considered frames. We notice that the evolution of the number of BBUs, over the day, follows the evolution of the hourly occupied GSM resources in Fig. 6.1. Interestingly, this shows that our strategy also allows to adapt the usage of network resources to the traffic evolution, suggesting important savings, in terms of computational resources, for both a Cloud-RAN and a Centralized-RAN architecture, and savings, in terms of energy, for a Centralized-RAN.

We present in Fig. 7.8 a subset of representative hourly geographical BBU-RRH association snapshots over the day. There, each RRH takes a color according to the BBU to which it is associated. The figure clearly shows how the geographical span and the size of clusters vary, over the day. At low traffic hours, e.g. 0:00 and 4:00, a few BBUs are enough to cover the whole city, with each covering a relatively large area. At high traffic hours, e.g. 8:00, 12:00 and 16:00, much more BBUs are needed in the network, with each handling the demand of several RRHs. Medium traffic hours present an intermediate behavior, as shown at 20:00. We remark that, for all traffic conditions, RRHs affected to the same BBU are generally closely located, as one would expect for the objective of reducing network handovers.

We take a closer look at the system dynamics at a finer temporal granularity, over a low traffic hour, so as to be able to track easily the evolution of clusters. We plot in Fig. 7.8 a subset of



(a) 0:00

(b) 4:00

(c) 8:00

(d) 12:00

(e) 16:00

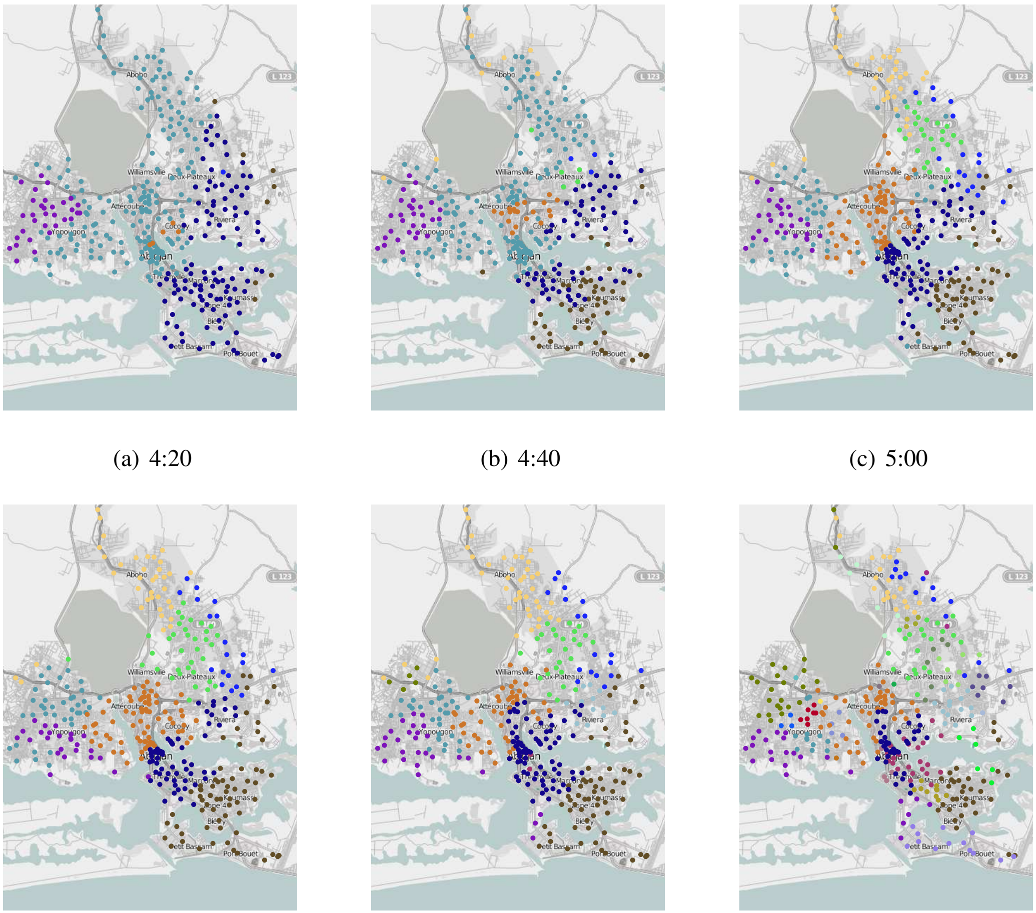
(f) 20:00

FIGURE 7.7: BBU-RRH associations over a representative set of snapshots over the whole day.

snapshots spanning between 4:20 and 6:00, with a difference of 20 minutes between a couple of successive snapshots. We notice that the structure of existing clusters smoothly changes as traffic evolves. We remark that, over some snapshots, it is possible for a cluster to be geographically split into two, such that the cluster does not form a connected component over the geographical space, at that time. Nevertheless, our clusters form connected components over the spatio-temporal space, meaning that, even if a cluster is split over a snapshot, it is actually connected over time.

7.6.4 Individual BBU behavior

We characterize the behavior of individual BBUs in Fig. 7.9 and Fig. 7.10, where we draw the distribution of number of RRHs and the number of calls per BBU, throughout the day. The candlesticks in the two figures, show the median, minimum and maximum values, together with the first and third quartile of the considered metric. Fig. 7.9 unveils a stable behavior, over day hours, in terms of number of RRHs per BBU, with only minor variations. Conversely, during



(a) 4:20

(b) 4:40

(c) 5:00

(d) 5:20

(e) 5:40

(f) 6:00

FIGURE 7.8: BBU-RRH associations over 6 consecutive snapshots.

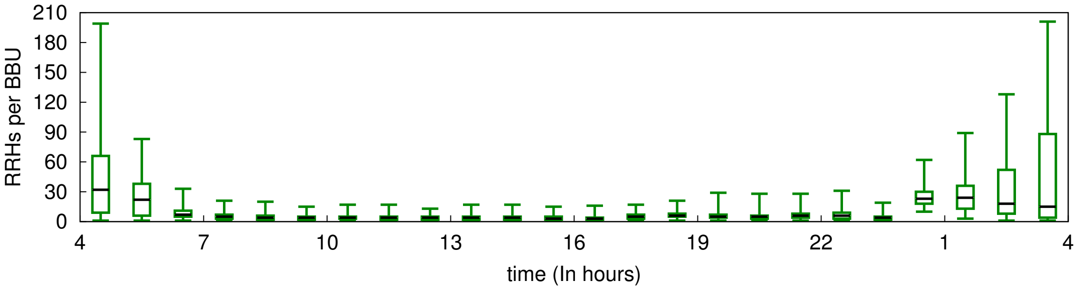


FIGURE 7.9: Distribution of the number of RRHs per BBU, over the day. The candlesticks show the median, minimum and maximum values, together with the first and third quartile for values obtained over each six hourly frames.

the night hours, we notice that the number of RRHs per BBU encounters significant variations, spanning from having one RRH per BBU, to cases where more than 200 RRHs are covered by the same BBU.

Still, these results do not provide information concerning the evolution of load, over individual BBUs. Fig. 7.10 allows to capture that, in terms of distribution of number of calls per BBU, over

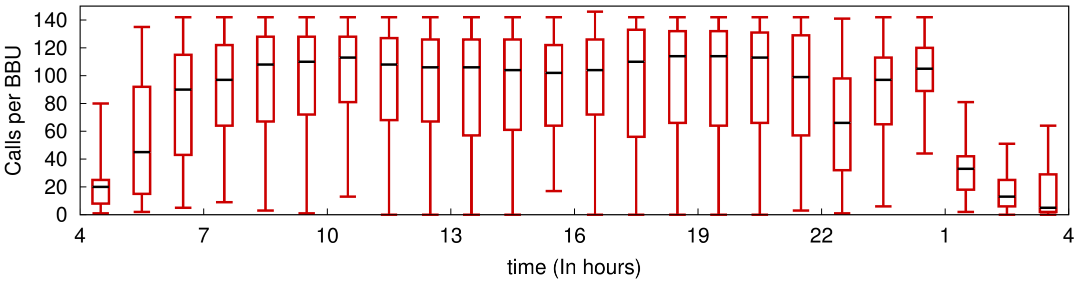


Figure 7.10: Distribution of the number of calls per BBU, over the day. The candlesticks show the median, minimum and maximum values, together with the first and third quartile for values obtained over each six hourly frames.

the day. We notice that the candlesticks follow the evolution of traffic, with more heterogeneous behaviors appearing at high traffic hours. Nevertheless, even at low traffic hours, one can still note important variations. Overall, the behavior of BBUs remains heterogeneous, indicating that, in a centralized architecture, an adequate dimensioning of the system can lead to important energy savings in the network.

7.6.5 Extension to other mobile technologies

Our results show that the proposed adaptive Cloud-RAN topology control scheme is beneficial for the overall system performance. However, the results that we derive are obtained based on GSM mobile traffic datasets. They thus reflect the benefits of virtualization techniques with respect to only one mobile technology, while Cloud-RAN is expected to encompass various standards.

Nevertheless, our main objective is to assess the capabilities of the our methodology and we expect to obtain important savings for other cellular network technologies. This is particularly true for cellular systems that support data communications, where a lot of applications run in the background and can imply much more handovers over the network than the case of GSM. In addition to that, LTE systems support high mobility speeds, translating into even more handovers on the network. This underlines the importance of our study for different mobile technologies. As part of our future work, we plan to extend our model to other cellular network standards, and consider datasets with voice, texting and data traffic activities.

7.7 Conclusion

In this chapter, we exploited mobile traffic datasets to explore the potential of a Cloud-RAN architecture for the management of users mobility. In particular, we evaluate possible savings in terms of handovers over the network resulting from a dynamic management of the Cloud-RAN topology. We use a time-varying graph model to study our problem, and employ a graph clustering strategy, allowing us to dynamically associate components of the Cloud-RAN architecture by accounting for both, users mobility and traffic demand.

We apply our strategy over a real-world mobile traffic dataset. Our results show that our method allows to reduce handovers in the network for a low and a high mobility scenarios by a percentage of almost 24%. Moreover, our strategy allows to adapt the system configuration to the traffic demand, suggesting promising supplementary energy and computational efforts savings in future Cloud-RAN and Centralized-RAN systems.

Part IV

Cellular Networks and Mobile Traffic Datasets: Conclusions and Perspectives

Chapter 8

Conclusions and open perspectives

8.1 Summary

The massive integration of mobile devices into our lives has raised a number of challenges for network operators. These devices are constantly connecting to the network with increasing traffic load, that network operators are hardly able to handle. However, as a result of their interaction with network infrastructure, they also produce logs, collected by network operators as mobile traffic datasets.

Mobile traffic datasets reflect human activity dynamics. Accordingly, they hold promising potential for future generations of cellular networks. They are especially useful for future dynamic networking solutions that adapt to traffic dynamics. Such possibilities require efforts oriented in two directions: analysis of mobile traffic datasets, so as to understand their main properties, and the design of adequate adaptive solutions, to improve the performance of the network. In this thesis, we conduct efforts with respect to these two points.

First, we target the analysis of mobile traffic datasets. We propose a framework that allows to **characterize large-scale datasets** and extract representative categories of consumption patterns. We obtain these categories by considering two complementary facets of mobile traffic: its volume and its distribution. The framework also allows to distinguish typical patterns from outlying ones in the network. We test our framework over real-world traffic datasets and show how it is capable of automatically grouping mobile traffic consumption patterns into a small set of significative categories and detecting untypical behaviors. Summarizing consumption patterns into a limited set of categories and characterizing the behavior of outlying events pave the way towards automatic management solutions of the cellular networks, with configurations that are suitable for each category.

Second, we exploit the potential of mobile traffic datasets for the evaluation of dynamic cellular network solutions. We tackle two problems of major importance for future 5G networks.

Namely, we study the problem of growing **energy consumption over cellular network infrastructure**, and propose a strategy capable of reducing it. The proposed strategy operates at the level of individual BSs. It dynamically adapts the power consumption of a BS to the evolution of traffic in its vicinity, i.e. its own traffic and that of its neighboring BSs, while ensuring a geographical coverage. By employing our strategy in a realistic large-scale environment, we show that savings of 27% can be achieved over a typical working day in an urban area.

Another problem that we consider is that of the **management of future Cloud-RAN**, with particular attention to users mobility. We propose a methodology that allows to dynamically adapt the network topology to the mobility of users, while accounting for their demand. We observe that savings of 24% can be achieved in a realistic urban large-scale environment.

As such, we shed light on two main topics for future cellular networks. We propose load-adaptive techniques that allow to cope with them and highlight the importance of mobile traffic datasets to evaluate them.

8.2 Open perspectives

The works presented in this thesis uncover the capabilities of mobile traffic datasets for the field of cellular networks. They also open up research directions for the future.

First, there is room for improvements in the methodologies that we propose. The capabilities of our mobile traffic characterization framework can be expanded by exploring different spatio-temporal granularities, considering various distance metrics and testing a variety of clustering strategies. All these aspects have a direct impact on the output of the framework and deserve further investigations.

Concerning our power reduction control scheme, it can also be improved. One can think of bringing ameliorations to its individual steps, e.g. including progressive power reductions instead of abrupt switch-off decisions for low traffic conditions. Besides, a quite interesting possibility is to employ advanced network optimization methods, capable of deriving optimal solutions for the complete problem.

Similarly, improvements can be brought to the methodology proposed for the dynamic configuration of Cloud-RAN topology over each step, e.g. by considering other clustering approaches or even employing optimization tools to infer optimal solutions.

Second, open questions relating to the practical implementation of these solutions remain to be addressed. This is especially critical for our networking solutions. Our power reduction control method needs to be complemented by cooperation schemes between BSs. These are needed to enable coordinations among BSs and prepare users handover procedures from one cell to the other as a result of power control decisions.

As for the Cloud-RAN topology control method, it is important to note that the derived configurations operate in an offline mode, i.e. by considering that we acquire a knowledge of future demands. Accordingly, operators may employ them over the network based on predictions of future traffic evolution. Another option would be to expand the scope of the method and modify it so as to react in real-time, with respect to current traffic state only. Both these solutions can be envisioned and further investigations are required to understand the potential and limits of each in terms of performance. Moreover, so far, we do not account for costs relating to virtualization, such as users migration costs from one virtual machine to the other, and their actual impact on the performance of the network.

Third, while in the thesis we separately consider the analysis and exploitation of mobile traffic datasets, blending them together, with automated procedures, can result in higher performance gains on one hand, and simpler network management procedures, on the other. As such, one can consider precise constraints of networking applications, and translate them into traffic patterns characteristics requirements for the framework. The latter would generate representative traffic patterns that account for them. Then, networking solutions would operate with a set of limited configurations, on top of the obtained traffic patterns, simplifying the management of the network.

Finally, the results obtained in this thesis are derived over one dataset, including voice traffic activity only. This goes along our objective to unveil the potential of such datasets and indicate the capabilities or proposed networking solutions. However, we plan to extend the work over other datasets in different contexts and with more diverse traffic activities that cover voice, messaging and data traffic. In fact, our traffic characterization framework has already been applied over a Telecom Italia dataset relating to activity of their users in Milan, underlining its capabilities to form again significant categories of traffic. Applications of networking solutions over other datasets and across different cellular network standards are planned for the future.

List of Publications

International Journal Articles

- D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, *Large-scale Mobile Traffic Analysis: a Survey*, IEEE Communications Surveys and Tutorials, to appear.

International Conference Articles

- M. Gramaglia, O. Trullols-Cruces, D. Naboulsi, M. Fiore, M. Calderon, *Vehicular Networks on Two Madrid Highways*, IEEE SECON 2014, Singapore, Singapore, June 2014.
- D. Naboulsi, R. Stanica, and M. Fiore. *Classifying Call Profiles in Large-scale Mobile Traffic Datasets*, IEEE INFOCOM 2014, Toronto, Canada, April 2014.
- D. Naboulsi, and M. Fiore. *On the Instantaneous Topology of a Large-scale Urban Vehicular Network: the Cologne case*, ACM MobiHoc 2013, Bangalore, India, August 2013.

Short Papers

- D. Naboulsi, M. Fiore, C.F. Chiasserini, *Assessment of Practical Energy Savings in Cellular Networks*, IEEE INFOCOM 2014 Student Workshop, Toronto, Canada, April 2014.
- D. Naboulsi, M. Fiore, R. Stanica, *On the Characterization of Mobile Calling Behaviors*, ACM MobiHoc 2013, Bangalore, India, August 2013.
- D. Naboulsi, M. Fiore, R. Stanica, *Human Mobility Flows in the City of Abidjan*, Netmob 2013, Boston, USA, May 2013.
- D. Naboulsi, M. Fiore, *The Connectivity of Cologne's Large-scale Vehicular Network*, IEEE INFOCOM 2013 Student Workshop, Torino, Italy, April 2013.

International Journal Articles - Under Revision

- D. Naboulsi, M. Fiore, *On the Instantaneous Topology of a Large-Scale Urban Vehicular Network: the Cologne Case.*

International Journal Articles - Under Review

- A. Furno, D. Naboulsi, R. Stanica, M. Fiore, *Mobile Demand Profiling for Cognitive Networking.*
- M. Gramaglia, O. Trullols-Cruces, D. Naboulsi, M. Fiore, M. Calderon, *Mobility and connectivity in highway vehicular networks: a case study in Madrid.*

Bibliography

[1] "Ericsson Mobility Report — On the Pulse of the Networked Society", Aug. 2014.

[2] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019", Feb. 2014.

[3] "http://www.o3bnetworks.com/growing-opportunities-in-rural-connectivity-across-all-markets/" (Last visited: 9/8/2015)

[4] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, J. Wang, "Large-Scale Measurement and Characterization of Cellular Machine-to-Machine Traffic", *IEEE/ACM Transactions on Networking*, 21(6):1960-1973, Dec. 2013.

[5] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, "Mobile Traffic Analysis: a Survey", *HAL-INRIA: Research Report, hal-01132385*, Mar. 2015.

[6] V.D. Blondel, A. Decuyper, G. Krings, "A Survey of Results on Mobile Phone Datasets Analysis", *arXiv:1502.03406 [physics.soc-ph]*, Feb. 2015.

[7] V. Soto, V. Frias-Martinez, J. Virseda, E. Frias-Martinez, "Prediction of Socioeconomic Levels Using Cell Phone Records", *UMAP*, Girona, Spain, Jul. 2011.

[8] A. Mehrotra, A. Nguyen, J. Blumenstock, V. Mohan, "Differences in Phone Use between Men and Women: Quantitative Evidence from Rwanda", *ICTD*, Atlanta, GE, USA, Mar. 2012.

[9] M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi, "Understanding Individual Human Mobility Patterns," *Nature*, 453(7196):779–782, Jun. 2008.

[10] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, "Ranges of Human Mobility in Los Angeles and New York," *IEEE PerCom Workshops*, Seattle, WA, USA, Mar. 2011.

[11] V. Soto, E. Frias-Martinez, "Automated Land Use Identification Using Cell-Phone Records", *ACM HotPlanet*, Washington, DC, USA, Jun. 2011.

[12] H. Zang, J. Bolot, “Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks”, *ACM MobiCom*, Montreal, Quebec, Canada, Sep. 2007.

[13] Z. Zhu, G. Cao, S. Zhu, S. Ranjan, A. Nucci, “A Social Network Based Patching Scheme for Worm Containment in Cellular Networks”, *IEEE Infocom*, Rio de Janeiro, Brazil, Apr. 2009.

[14] Nokia Mobile Data Challenge: http://research.nokia.com/page/12000 (Last visited: 9/8/2015)

[15] Orange Data for Development Challenge: http://www.d4d.orange.com (Last visited: 9/8/2015)

[16] Telecom Italia Big Data Challenge: http://www.telecomitalia.com/tit/en/bigdatachallenge (Last visited: 9/8/2015)

[17] M. Kafsi, E. Kazemi, L. Maystre, L. Yartseva, M. Grossglauser, P. Thiran, “Mitigating Epidemics through Mobile Micro-Measures”, *NetMob*, Boston, MA, USA, May 2013.

[18] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, C. Ziemlicki, “Data for Development: the D4D Challenge on Mobile Phone Data”, *arXiv: 1210.0137 [cs.CY]*, Sep. 2012.

[19] S. Hoteit, S. Secci, G. Pujolle, Z. He, C. Ziemlicki, Z. Smoreda, C. Ratti, “Content Consumption Cartography of the Paris Urban Region Using Cellular Probe Data”, *UrbaNe*, Nice, France, Dec. 2012.

[20] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, J. Wang, “Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network”, *IEEE Infocom*, Orlando, Florida, USA, Mar. 2012.

[21] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi, “Limits of Predictability in Human Mobility,” *Science*, 327(5968):1018–1021, Jan. 2010.

[22] C.-P. Wei, I. T. Chiu, “Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach”, *Expert systems with applications* 23(2): 103-112, 2002.

[23] J. Guo, F. Liu, Z. Zhu, “Estimate the Call Duration Distribution Parameters in GSM System Based on K-L Divergence Method”, *IEEE WiCom*, Shanghai, China, Sep. 2007.

[24] Cost 231 Final Report, http://www.lx.it.pt/cost231/ (Last visited: 9/8/2015)

[25] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, “Numerical Recipes Third Edition: the Art of Scientific Computing”, *Cambridge university press*, 2007.

[26] C. Williamson, E. Halepovic, H. Sun, Y. Wu, “Characterization of CDMA2000 Cellular Data Network Traffic”, *IEEE LCN*, Sydney, Australia, Nov. 2005.

[27] R. Keralapura, A. Nucci, Z. L. Zhang, L. Gao, “Profiling Users in a 3G Network Using Hourglass Co-clustering”, *ACM MobiCom*, Chicago, Illinois, USA, Sep. 2010.

[28] U. Paul, A. P. Subramanian, M. M. Buddhikot, S. R. Das, “Understanding Traffic Dynamics in Cellular Data Networks”, *IEEE Infocom*, Shanghai, PRC, Apr. 2011.

[29] M. Z. Shafiq, L. Ji, A. X. Liu, J. Wang, “Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices”, *ACM SIGMETRICS*, San Jose, California, USA, Jun. 2011.

[30] Y. Zhang, A. Arvidsson, “Understanding the Characteristics of Cellular Data Traffic”, *ACM SIGCOMM CellNet Workshop*, Helsinki, Finland, Aug. 2012.

[31] E. Mucelli, A. C. Viana, K. P. Naveen, C. Sarraute, “Measurement-driven Mobile Data Traffic Modeling in a Large Metropolitan Area”, *IEEE PerCom*, St. Louis, MO, USA, Mar. 2015.

[32] Y. Wang, M. Faloutsos, H. Zang, “On the Usage Patterns of Multimodal Communication: Countries and Evolution”, *IEEE GI*, Turin, Italy, Apr. 2013.

[33] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, C. Ratti, “Towards Estimating the Presence of Visitors from the Aggregate Mobile Phone Network Activity They Generate”, *CUPUM*, Hong Kong, PRC, Jun. 2009.

[34] H. Hohwald, E. Frias-Martinez, N. Oliver, “User Modeling for Telecommunication Applications: Experiences and Practical Implications”, *UMAP*, Big Island, Hawaii, USA, Jun. 2010.

[35] J. C. Cardona, R. Stanojevic, N. Laoutaris, “Collaborative Consumption for Mobile Broadband: A Quantitative Study”, *ACM CoNext*, Syndney, Australia, Dec. 2014.

[36] R. M. Pulselli, P. Romano, C. Ratti, E. Tiezzi, “Computing Urban Mobile Landscapes Through Monitoring Population Density Based on Cell-Phone Chatting”, *International Journal of Design and Nature and Ecodynamics*, 3(2): 121-134, 2008.

[37] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, J. Blat, “Digital Footprinting: Uncovering Tourists with User-Generated Content”, *IEEE Pervasive Computing*, 7(4):36–43, Oct. 2008.

[38] C. Ratti, R. M. Pulselli, S. Williams, D. Frenchman, “Mobile Landscapes: Using Location Data from Cell-Phones for Urban Analysis”, *Environment and Planning B Planning and Design* 33(5): 727, 2006.

[39] D. Willkomm, S. Machiraju, J. Bolot, A. Wolisz "Primary Users in Cellular Networks: A Large-Scale Measurement Study", *IEEE DySPAN*, Chicago, Illinois, USA, Oct. 2008.

[40] B. Csáji, A. Browet, V. A. Traag, J. C. Delvenne, E. Huens, P. V. Dooren, Z. Smoreda, V.D. Blondel, "Exploring the Mobility of Mobile Phone Users", *Physica A*, 392(6):1459–1473, Jun. 2013.

[41] M. Cerinsek, J. Bodlaj, V. Batagelj, "Symbolic Clustering of Users and Antennae", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[42] S. Almeida, J. Queijo, L. M. Correia, "Spatial and Temporal Traffic Distribution Models for GSM", *IEEE VTC Fall*, Amsterdam, Netherlands, Sep. 1999.

[43] B. Cici, M. Gjoka, A. Markopoulou, and C. T. T Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology" *ACM MobiHoc*, Hangzhou, China, Jun. 2015.

[44] M. R. Vieira, V. Frias-Martinez, N. Oliver, E. Frias-Martinez, "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics", *IEEE SocialCom*, Minneapolis, Minnesota, USA, Aug. 2010.

[45] I. Trestian, S. Ranjan, A. Kuzmanovic, A. Nucci, "Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network", *ACM IMC*, Chicago, IL, USA, Nov. 2009.

[46] R. Trasarti, A. M. Olteanu-Raimond, M. Nanni, T. Couronne, B. Furletti, F. Giannotti, Z. Smoreda, C. Ziemlicki "Discovering Urban and Country Dynamics from Mobile Phone Data with Spatial Correlation Patterns", *NetMob*, Boston, MA, USA, May 2013.

[47] F. H. Z. Xavier, L. M. Silveira, J. M. Almeida, A. Ziviani, C. H. S. Malab, H. M. Neto, "Analyzing the Workload Dynamics of a Mobile Phone Network in Large Scale Events", *UrbaNe*, Nice, France, Dec. 2012.

[48] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, J. Wang, "A First Look at Cellular Network Performance During Crowded Events", *ACM SIGMETRICS*, Pittsburgh, PA, USA, Jun. 2013.

[49] P. Paraskevopoulos, T. C. Dinh, Z. Dashdorj, T. Palpanas, L. Serafini, "Identification and Characterization of Human Behavior Patterns from Mobile Phone Data", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[50] D. M. Gowan, N. Hurley, "Regional Development - Capturing a Nation's Sporting Interest Through Call Detail Analysis", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[51] F. H. Z. Xavier, L. M. Silveira, J. M. Almeida, C. H. S. Malab, A. Ziviani, H. T. Marques-Neto, "Understanding Human Mobility Due to Large-Scale Events", *NetMob*, Boston, MA, USA, May 2013.

[52] D. Pastor-Escuredo, T. Savy, M. A. Luengo-Oroz, "Can Fires, Night Lights, and Mobile Phones Reveal Behavioral Fingerprints Useful for Development?", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[53] S. V. D. Elzen, J. Blaas, D. Holten, J. K. Buenen, J. J. V. Wijk, R. Spousta, A. Miao, S. Sala, S. Chan, "Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics Approach", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[54] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, A. L. Barabasi "Uncovering Individual and Collective Human Dynamics from Mobile Phone Records", *Journal of Physics A: Mathematical and Theoretical* 41(22): 224015, 2008.

[55] M.F. Dixon, S. P. Aiello, F. Fapohunda, W. Goldstein, "Detecting Mobility Patterns in Mobile Phone Data from the Ivory Coast", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[56] nanavati06 M. Karsai, N. Perra, A. Vespignani, "Time Varying Networks and the Weakness of Strong Ties", *Scientific Reports*, 4(4001):1-7, Feb. 2014.

[57] A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasagupta, S. Mukherjea, A. Joshi, "On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications", *ACM CIKM*, Arlington, VA, USA, Nov. 2006.

[58] Y. Altshuler, M. Fire, E. Shmueli, Y. Elovici, A. Bruckstein, A. S. Pentland, D. Lazer, "The Social Amplifier – Reaction of Human Communities to Emergencies", *Journal of Statistical Physics*, 152(3): 399-418, Aug. 2013.

[59] B. Zong, P. Bogdanov, A. K. Singh, "Constrained Link Prediction on the D4D Dataset", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[60] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, J. Leskovec, "Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions", *ACM KDD*, Las Vegas, NV, USA, Aug. 2008.

[61] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, M. A. de Menezes, K. Kaski, A.-L. Barabasi, J. Kertesz, "Analysis of a Large-Scale Weighted Network of One-to-One Human Communication", *New Journal of Physics*, 9(179):1-27, Jun. 2007.

[62] D. Doran, V. Mendiratta, C. Phadke, H. Uzunalioglu, "The Importance of Outlier Relationships in Mobile Call Graphs", *IEEE ICMLA*, Boca Raton, FL, USA, Dec. 2012.

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

[64] P. Sneath, “The Application of Computers to Taxonomy”, *Journal of General Microbiology*, 17(1): 201–226, 1957.

[65] P. Sneath, R. Sokal, “Unweighted Pair Group Method with Arithmetic Mean”, *Numerical Taxonomy*, 230–234, 1973.

[66] J. H. Ward, “Hierarchical Grouping to Optimize an Objective Function”, *Journal of the American Statistical Association*, 58(301): 236–244, 1963.

[67] L. Kaufman, P. Rousseeuw, “Finding Groups in Data: An Introduction to Cluster Analysis”, *John Wiley & Sons*, 34, 1990.

[68] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations”, Berkeley Symposium on Mathematical Statistics and Probability, 1(14): 281–297, 1967.

[69] G. Ball, D. Hall, “A clustering technique for summarizing multivariate data”, *Behavioral Science*, 12(2): 153–155, 1967.

[70] Q. Lin, “Mobile Customer Clustering Analysis based on Call Detail Records”, *Communications of the IIMA* 7(4): 95-100, 2007.

[71] R. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, C. Volinsky, “Clustering Anonymized Mobile Call Detail Records to Find Usage Groups”, *PURBA*, San Francisco, CA, USA, Jun. 2011.

[72] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, “Social Ties and their Relevance to Churn in Mobile Telecom Networks”, *EDBT*, Nantes, France, Mar. 2008.

[73] F. Ben Abdesslem, A. Lindgren, “Large Scale Characterisation of YouTube Requests in a Cellular Network”, *IEEE WoWMoM*, Sydney, Australia, Mar. 2014.

[74] T. Couronné, Z. Smoreda, A. M. Olteanu, “Chatty Mobiles: Individual Mobility and Communication Patterns”, *NetMob*, Boston, MA, USA, Oct. 2011.

[75] I. T. Jollifee, “Principal component analysis”, *John Wiley & Sons*, Ltd, 2002.

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

[77] 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.

[78] R. Belo, P. Ferreira, “Is Social Influence Always Positive? Evidence from a Very Large Mobile Network”, *Economics of Information Technology and Digitization Workshop*, Boston, MA, USA, Jul. 2013.

[79] G. Szabo, A. L. Barabasi, “Network Effects in Service Usage”, *arXiv pre-print*, arXiv:physics0611177, Nov. 2006.

[80] Q. Xu, A. Gerber, Z. M. Mao, J. Pang, “AccuLoc: Practical localization of Performance Measurement in 3G Networks”, *ACM MobiSys*, Washington, DC, USA, Jun. 2011.

[81] A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, J. Erman, “To Cache or Not to Cache: The 3G Case”, *IEEE Internet Computing*, 15(2):27–34, Mar. 2011.

[82] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, V. Erramilli, Y. Grunenberger, “Is There a Case for Mobile Phone Content Pre-Staging?”, *ACM CoNEXT*, Santa Barbara, CA, USA, Dec. 2013.

[83] P. Wang, M. C. Gonzalez, C. A. Hidalgo, A. L. Barabasi, “Understanding the Spreading Patterns of Mobile Phone Viruses”, *Science 324*, no. 5930: 1071-1076, 2009.

[84] R. Agarwal, V. Gauthier , M. Becker, “Information Dissemination Using Human Mobility in Realistic Environment - (E-Inspire)”, *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[85] Y. Zhu, C. Zhang, Y. Wang, “Mobile Data Delivery Through Opportunistic Communications Among Cellular Users: A Case Study for the D4D Challenge”, *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[86] F. Yu, G. Xue, H. Zhu, Z. Hu, M. Li, G. Zhang, “Cutting without Pain: Mitigating 3G Radio Tail Effect on Smartphones”, *IEEE Infocom*, Turin, Italy, Apr. 2013.

[87] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, H. Yan, “Modeling Web Quality-of-Experience on Cellular Networks”, *ACM MobiCom*, Maui, Hawaii, USA, Sep. 2014.

[88] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, J. Wang, “Understanding the Impact of Network Dynamics on Mobile Video User Engagement”, *ACM SIGMETRICS*, Austin, TX, USA, Jun. 2014.

[89] H. Zang, J. Bolot, “Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study”, *ACM MobiCom*, Las Vegas, NV, USA, Sep. 2011.

[90] Y. A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, “Unique in the Crowd: The Privacy Bounds of Human Mobility”, *Scientific Reports* 3, 2013.

[91] Y. Song, D. Dahlmeier, S. Bressan, "Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data", *ACM PIR*, Queensland, Australia, Jul. 2014.

[92] G. Acs, C. Castelluccia, "A Case Study: Privacy Preserving Release of Spatio-Temporal Density in Paris", *ACM SIGKDD*, New York, NY, USA, Aug. 2014.

[93] http://www.anfr.fr/fr/observatoire-deploiement-2g3g4g/les-resultats-de-lobservatoire/juin-2015.html (Last visited: 9/8/2015)

[94] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C.-X. Wang, T. A. Le, M. R. Nakhai, J. Zhang, and L. Hanzo, "Green Radio: Radio Techniques to Enable Energy-Efficient Wireless Networks", *IEEE Communications Magazine*, 49(6): 46 - 54, May. 2011.

[95] Z. Hasan, H. Boostanimehr, V. K. Bhargava, "Green Cellular Networks: A survey, some Research Issues and Challenges", *IEEE Communications Surveys & Tutorials*, 13(4):524-540, Nov. 2011.

[96] L. M. Correia, D. Zeller, O. Blume, Y. Jading, I. Godor, G. Auer, L. V. D. Perre, "Challenges and Enabling Technologies for Energy Aware Mobile Radio Networks", *IEEE Communcations Magazine*, 48(11): 66-72, Nov. 2010.

[97] A. D. Domenico, E. C. Strinati, A. Capone, "Enabling Green Cellular Networks: A Survey and Outlook", *Computer Communications*, 37: 5-24, Jan. 2014.

[98] M. A. Marsan, L. Chiaraviglio, D. Ciullo, M. Meo, "Optimal Energy Savings in Cellular Access Networks" , *IEEE Communications Workshop*, Dresden, Germany, Jun. 2009.

[99] L. Chiariviglio, D. Ciullo, M. Meo, M. A. Marsan, "Energy-aware UMTS Access Networks", *IEEE WPMC*, Lapland, Finland, Sep. 2008.

[100] J.M. Kelif, M. Coupechoux, F. Marache, "Limiting Power Transmission of Green Cellular Networks: Impact on Coverage and Capacity", *IEEE ICC*, Cape Town, South Africa, May 2010.

[101] M. A. Marsan, M. Meo, "Energy Efficient Management of Two Cellular Access Networks", *ACM SIGMETRICS Performance Evaluation Review*, 37(4): 69-73, Mar. 2010.

[102] K. Dufková, M. Bjelica, B. Moon, L. Kencl, J.-Y. Le Boudec, "Energy Savings for Cellular Network with Evaluation of Impact on Data Traffic Performance", *IEEE EW*, Lucca, Italy, Apr. 2010.

[103] E. Oh, K. Son, B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks", *IEEE Transactions on wireless communications*, 12(5): 2126-2136, May 2013.

[104] G. Micallef, P. Mogensen, H.-O. Scheck, “Cell Size Breathing and Possibilities to Introduce Cell Sleep Mode”, *IEEE EW*, Lucca, Italy, Apr. 2010.

[105] Z. Niu, Y. Wu, J. Gong, Z. Yang, “Cell Zooming for Cost-Efficient Green Cellular Networks”, *IEEE Communications Magazine*, 48(11): 74-79, Nov. 2010.

[106] C. Liu, B. Natarajan, H. Xia, “Small Cell Base Station Sleep Strategies for Energy Efficiency”, *IEEE Transactions on Vehicular Technology*, 99, Mar. 2015.

[107] D. Tipper, A. Rezgui, P. Krishnamurthy, P. Pacharintanakul, “Dimming Cellular Networks”, *IEEE Globecom*, Miami, Florida, USA, Dec. 2010.

[108] R. Combes, S. E. Elayoubi, A. Ali, L. Saker, T. Chahed, “Optimal Online Control for Sleep Mode in Green Base Stations”, *Computer Networks*, 78(26): 140-151, Feb. 2015.

[109] C. Peng, S. B. Lee, S. Lu, H. Luo, H. Li, “Traffic-driven Power Saving in Operational 3G Cellular Networks”, *ACM Mobicom*, Las Vegas, Nevada, USA, Sep. 2011.

[110] T. Han, N. Ansari, “On Optimizing Green Energy Utilization for Cellular Networks with Hybrid Energy Supplies”, *IEEE Transactions on Wireless Communications*, 12(8): 3872-3882, Aug. 2013.

[111] M. A. Marsan, G. Bucalo, A. D. Caro, M. Meo, Y. Zhang, “Towards Zero Grid Electricity Networking: Powering BSs with Renewable Energy Sources”, *Italian Networking Workshop*, Bormio, Italy, Jan. 2013.

[112] O. Arnold, F. Richter, G. Fettweis, O. Blume, “Power consumption modeling of different BS types in heterogeneous cellular networks.” IEEE Future Network and Mobile Summit, June 2010.

[113] “Digital Cellular Telecommunication System (phase 2)”; Radio Transmission and Reception, ETSI, 3GPP TS 05.05, Ver 8.20.0, Release 1999.

[114] China Mobile Research Institute, “C-RAN the road towards green ran”, *Technical Report*, Beijing, China, Oct. 2011.

[115] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, L. Dittmann, “Cloud RAN for Mobile Networks—A Technology Overview”, *IEEE Communication Surveys and Tutorials*, 17(1): 405-426, Jan. 2015.

[116] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, Y. Kishi, “Colony-RAN architecture for future cellular network”, *IEEE FutureNetw*, Berlin, Germany, Jul. 2012.

[117] A. Checko, H. L. Christiansen, M. S. Berger, “Evaluation of energy and cost savings in mobile Cloud RAN”, *Opnetwork*, Washington, DC, USA, 2013.

[118] A. Checko, H. Holm, H. Christiansen, “Optimizing small cell deployment by the use of C-RANs”, *EW*, Barcelona, Spain, May 2014.

[119] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, G.-K. Chang, “The case for reconfigurable backhaul in cloud-RAN based small cell networks”, *IEEE Infocom*, Turin, Italy, Apr. 2013.

[120] M. Madhavan, P. Gupta, M. Chetlur, “Quantifying Multiplexing Gains in a Wireless Network Cloud”, *IEEE ICC SA-AN*, Ottawa, Canada, Jun. 2012.

[121] T. Werthmann, H. G.-Lipski, M. Proebster, “Multiplexing Gains Achieved in Pools of Baseband Computation Units in 4g Cellular Networks”, *IEEE PIMRC*, London, UK, Sep. 2013.

[122] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, T. Woo, “CloudIQ: a framework for processing base stations in a data center”, *IEEE MobiCom*, Istanbul, Turkey, Aug. 2012.

[123] L. Liu, F. Yang, R. Wang, Z. Shi, A. Stidwell, D. Gu, “Analysis of Handover Performance Improvement in Cloud-RAN Architecture”, *IEEE ICST CHINACOM*, Kunming, China, Aug. 2012.

[124] S. Namba, T. Warabino, S. Kaneko, “BBU-RRH switching schemes for centralized RAN”, *IEEE ICST CHINACOM*, Kunming, China, Aug. 2012.

[125] D. Zhu, M. Lei, “Traffic and Interference-aware Dynamic BBU-RRU Mapping in C-RAN TDD with Cross-subframe Coordinated Scheduling/Beamforming”, *IEEE ICC OWITSN*, Budapest, Hungary, Jun. 2013.

[126] K. Wang, M. Zhao, W. Zhou, “Graph-Based Dynamic Frequency Reuse in Cloud-RAN”, *IEEE WCNC*, Istanbul, Turkey, Apr. 2014.

[127] W. Klotz, “Graph Coloring Algorithms”, *Mathematics Report*, 5: 1-9, 2002.

[128] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, “Fast unfolding of communities in large networks”, *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, Oct. 2008.

[129] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, D. Wagner, “Maximizing Modularity is hard”, *arXiv: 0608255 [physics.data-an]*, Aug. 2006.

[130] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, O. Spatscheck, “Characterizing Radio Resource Allocation for 3G Networks”, *ACM IMC*, Melbourne, Australia, Nov. 2010.