From Networks to Data and Back Again

Razvan Stanica

N° Identificateur

HABILITATION À DIRIGER DES RECHERCHES

présentée devant

l'Institut National des Sciences Appliquées de Lyon
et l'Université Claude Bernard LYON 1

From Networks to Data and Back Again

A Story of Wireless Networks in the 21st Century

Spécialité : Informatique

par

Razvan Stanica

Soutenue le 21/11/2019 devant la Commission d'examen

Andrzej DUDA	Professeur	Grenoble INP	Rapporteur
Xavier LAGRANGE	Professeur	IMT Atlantique	Rapportuer
Mahesh K. MARINA	Professeur	Univ. of Edinburgh	Rapporteur
André-Luc BEYLOT	Professeur	INP Toulouse	Examinateur
Isabelle GUÉRIN LASSOUS	Professeur	Univ. Lyon 1	Examinateur
Catherine ROSENBERG	Professeur	Univ. of Waterloo	Examinateur
Fabrice VALOIS	Professeur	INSA Lyon	Examinateur

CITI Lab - Centre d'Innovation en Télécommunications et Intégration des Services

Acknowledgements

A research journey is not a solitary one, and several persons deserve to be thanked right from the beginning. First of all, my work during the last seven years has been shaped by three role models, colleagues and friends. So I would like to thank Fabrice, for continuously challenging me, for his support and advice, and for always being able to make me laugh. This manuscript would have never been possible without his insistence and, at the same time, long patience. I would also like to thank Hervé for trusting me right from the beginning (or at least for giving me this feeling). He knows better than anyone how to manage my ranting and how to keep me (relatively) focused. Finally, I have to thank Marco, for our long scientific and non-scientific talks, for countless beers and for bringing me in the football team. My scientific evolution would have been very different without him.

I am also thankful to those who guided my first steps in the research world, in Toulouse. André-Luc is the one who recruited me and taught me what research should be like. He will always be my model for scientific quality and human integrity (although I will never program in Fortran). Manu was always there to help me formulate the right questions, while also sharing his own research problems (which strangely often involved goats). He is probably the most talented teacher I ever had and my passion for networks is largely a consequence of his classes.

I have the chance to defend this work in front of a jury formed of people I admire and appreciate deeply. I feel very honored they accepted to review this manuscript and I would like to thank them. The first things I learned about mobile networks came from a red book about GSM, authored by Xavier. His books continued to be my reference and I was very proud (and scared) when he accepted to review my work, so I would like to thank him for this. To this day, Andrzej's work on IdleSense remains one of the best papers I ever read. I followed his work and read his papers since I was in my first year as a PhD student, so having him accept with so much kindness to review my manuscript

is an immense pleasure. At the end of my PhD thesis, I put up a list of the 10 research groups I would like to join and contacted the professors in lead of these groups. One of those professors was Mahesh (who probably does not remember this). He promptly and kindly answered at the time, and so he did now, when I contacted him to review this manuscript. Although we never properly collaborated, Isabelle was one of the first people I met when I moved to Lyon. She actually called me 2h after moving from Toulouse asking me to submit the final version of an invited article at a conference she was organizing (which I did from a McDonald's, since I had no Internet connection at my place). Ever since, I take every opportunity to infiltrate her busy schedule and learn from her. Finally, I would like to thank Catherine for managing to include this event in her agenda (I know it was difficult). The time I spent at Waterloo and our discussions shaped my research during the last three years. I am still shocked and in total disbelief when she calls *me* with technical or scientific questions, and my biggest regret is that I do not have more time to dedicate to our collaboration.

Of course, I want to thank all the students and post-docs I collaborated with. I learned a lot from each of them, even though some will not feature in the story told by this manuscript. So, my thoughts go to, in alphabetical order: Abderrahman, Alexis, Angelo, Assia, Diala, Domga, Elli, Ines, Irfan, Jad, Patrice, Romain, Silvia, Solo, Yota, Zung. I would also like to thank my colleagues from the Urbanet/Agora team, especially Oana, Walid, Khaled and Isabelle. I could not have wished for a better company every day. And I would like to thank everyone at the CITI lab, where I could always find interesting discussions or funny chat, depending on what I needed.

I must admit that this manuscript has been written during evenings, nights, weekends and holidays. So many thanks to those who were supposed to benefit from this time, my friends and family, who already see me quite rarely and now saw themselves neglected for a weird academic work. A special thank you goes to Oana, who supported me through this as she supported me through everything for the last 15 years, with a lot of patience and quite a few laughs. Without her idea of renting a pool house and feeding me with crackers and wine, this manuscript would never have been finished.

Finally, I would like to thank all my former and current students at the Telecommunications department. They challenge me every day and I am always amazed to see how much they mature in only three years. And I would also like to thank all my teachers, from primary school to university, in Romania

and in France. I learned something from each and every one of them and I put something from what they though me in everything I do.

Table of contents

Chapter 1

A Warm Welcome

Usually denoted as Introduction

> Far over the misty mountains cold
> To dungeons deep and caverns old
> We must away ere break of day
> To seek the pale enchanted gold.
>
> ---
>
> *The Hobbit, or, There and Back Again*
> J.R.R. Tolkien (1937)

Despite the numerous references to *There and Back Again*, this document is not about hobbits, nor is it as well written as Tolkien's tales. However, the story behind this manuscript resembles somehow to Bilbo Baggins' adventures. First of all, just like for the famous hobbit, the work I discuss here involved an actual geographical displacement on my side. Born and raised in southwest Romania, my graduate studies took me (entirely by chance) to Toulouse, a trip of around 2000 km. There, mentored by a pair of Gandalfs (without the beards though), I learned about how computer networks interconnect the entire world. Particularly interested by the wireless technologies that allow us users to see to our daily routines while remaining connected to everyone and everything, I remained in Toulouse for doctoral studies on this topic. More precisely, my PhD thesis, defended in November 2011, contributed to the field of vehicular networks, proposing mechanisms capable of coping with congestion at the medium access control layer. Adding some other 500 km, my journey continued in September 2012, when I joined INSA Lyon, the CITI laboratory and what was at the time the recently proposed Inria Urbanet team. This is where I met several companions, who facilitated, steered and assisted my research for the last seven

years. This document, summarizing the *lyonnaise* stage of my trip, also tells their stories.

The pages you will read also draw a second parallel with the fantastic quest from Tolkien's book, as they describe my exploration of new research subjects. I arrived in Lyon as a young researcher with expertise in wireless networks, but I soon started flirting with different other topics, such as data analysis, machine learning or user privacy. This exploration resulted in unexpected meetings and collaborations with fellow researchers from fields such as transportation, economics, urban planning, psychology or political sciences. While this multi-disciplinary experience clearly contributed to shape me as a scientist, my journey did not seem complete until I finally came back and applied these newly learned concepts to my *home* field of mobile networks. Now that the loop is closed, it is time to begin our story.

1.1 Research topics

Most of those who will read this document probably agree with my opinion that computer networks is the best research topic one could take interest in. It could therefore seem pointless to have here a historical discussion of the field, likely known to all the readers. But the way we interpret historical events is subjective and the lessons we learn from them are personal. This means that, in order to understand my vision of research challenges in the field, one needs first to understand the way I see its history.

Communication is an innate human need and people have continuously imagined new ways of extending their communication range, from visual signals to the use of animals to carry messages. Communication technologies as known today have their origins in the electrical engineering field at the end of the 19th century. The efforts at that time were mostly focused on designing the end-user devices, such as telephones, radio towers or radio receivers. The network connecting these devices was a basic one, the only intermediary devices being human handled switches.

With the evolution of electronics, human operators disappeared and digital core networks appeared. The early networks were characterized by a mono-application vision: telephony networks, radio broadcast networks, television broadcast networks, and so on. Even the first computer networks were mono-application, connecting terminals and mainframes or computer and printers.

This changed with the beginning of the Internet and it was formalized in the proposal of the layered networking approach that still stands today. While multiple applications were now allowed to share the same network, the traffic they generated was still quite similar in terms of service requirements, and the network was treating everything in a *best effort* way. Moreover, the end devices were at the time quite homogeneous: the Internet was interconnecting large computers from research institutes and universities. These devices were already a technological miracle and very expensive. The networking devices interconnecting them were relatively cheaper, with low processing power and memory. The limits of early hubs, switches and routers shaped the protocols used by the network: the Internet Protocol (IP) is as basic as it can get, and the Transport Control Protocol (TCP) only runs on the end devices. In fact, while the *end-to-end argument* might have become a philosophical principle in the last two decades, its origins are in the enormous differences between the capabilities of end devices and those of networking devices, which made it logical to place most of the computing-heavy network functions on the end devices.

Traffic differentiation and quality of service (QoS) have slowly but surely evolved in the networking world. While QoS-oriented technologies, such as Asynchronous Transfer Mode (ATM), HiperLAN or WiMax failed in the market, the concept of treating traffic differently depending on application requirements is now well established. If anything, the problem is to choose between the multitude of available solutions: Integrated Services (IntServ) and Differentiated Services (DiffServ) in the IP world, IEEE 802.1p and IEEE 802.11e at the data-link layer, bearers and the much touted slices in mobile networks, all try to provide differentiated QoS.

The fact that IP packets generated by a voice over IP (VoIP) application and those coming from a file transfer application should be treated in a different way by the network probably seems obvious to all the readers. However, as I found out while exploring non-networking problems, this is not so obvious to application developers or to the rising class of *privacy* experts. Indeed, applications deployed today never set up the proper bits in the IP headers, as required by DiffServ, or they never select anything else than the default bearer in a mobile operating system. Whether this is because developers lack networking skills, or because the network does not properly expose these options through an application programming interface (API), it is less important. But the result is that, for the last two decades, QoS classification has been done in the network, by using deep packet inspection (DPI), and the traffic is being divided in classes by the

network operator, not by the application developers. The resulting computer security risks and the increase in privacy awareness produced more and more applications with end-to-end encrypted traffic. As a consequence, the state of QoS differentiation in 2018 is that operators mainly distinguish two traffic classes in their networks: VoIP and everything else.

Meanwhile, end devices and network devices also evolved. With the success of handheld smartphones and tablets, not to mention the much announced revolution of the Internet of Things (IoT), a significant heterogeneity can be observed on this side, as these small devices now connect to powerful servers somewhere on the Internet. Moreover, some end devices today have even less resources than the networking devices. Indeed, modern routers are resourceful and expensive machines, capable of much faster processing than most end devices. To complete the picture, a series of devices denoted as *middleboxes* were slowly added inside the networks, to enable operations such as traffic filtering, network address translation, TCP optimization, or load balancing, to name just a few. Of course, relationships exist among all these functions implemented by middleboxes, which can easily lead to networking configuration errors: the story of a major operator who was shaping the network traffic output by a TCP optimizer is a well-known anecdote. Correctly configuring a network therefore becomes a complex task, especially as we evolve towards software defined networking (SDN), where the different network functions are virtualized and can be dynamically installed and run on generic hardware.

This summarizes the state of the networking nation at the time when my research takes place. Freshly out from a PhD thesis on vehicular networks, where extremely urgent safety messages have to coexist with best effort comfort applications, I was already convinced that networking mechanisms have to be traffic-aware. The obvious part, that I was not aware of, but which I discovered very quickly, was that traffic itself was produced by users, and it therefore depended on their habits, mobility patterns and preferences. My collaborations with colleagues from social sciences can be explained by this common interest: while they are trying to understand human behavior, I am trying to understand network users. Such collaborations between social and computer scientists are not only a mutualization of knowledge, but also an enabler of new tools and methodologies. For many years, social sciences were based on surveys and qualitative studies conducted on a limited number of subjects. Since communication devices are now an integral part of our daily lives, they allow collecting accurate and massive data regarding the user actions, meaning that

computer networks research can not only benefit from social sciences insight on human behavior, but also assist social scientists to expand their knowledge. Therefore, my research work gradually followed three intertwined paths: *i)* collecting data using network devices, *ii)* analyzing this data to detect patterns, classes and anomalies, and *iii)* conceive and evaluate traffic-aware networking solutions. In this document, I will try to convince the reader that, now that network devices, storage and computing resources allow it, the way we do networking must evolve: network devices must continuously monitor network traffic, analyze it and forecast its behavior, providing this information to traffic-aware network mechanisms which can reconfigure and adapt the network in real time.

For full disclosure, a small part of my work during the seven last years did not respect the above classification and it will not be addressed in this document. First of all, I continued contributing to the vehicular networks field, probably less than what would I have hoped for[1]. Second, I am still pursuing, slowly but surely, another interest developed during my PhD student years: the analytical and simulation modeling of medium access control (MAC) solutions in wireless networks, either Aloha or carrier-sense based. Finally, a part of my research was dedicated to the study of wireless networks with a strong asymmetry between the uplink and the downlink. This asymmetry, which can appear in both directions[2], challenges many well-established ideas in the networking community and generates a very interesting research playground.

1.2 Document structure

This document summarizes research I conducted since September 2012 at the CITI laboratory, jointly affiliated to INSA Lyon and Inria Rhone-Alpes. During this time, I was part of the Inria Urbanet team, which became the Inria Agora team in 2017. Most of the results I will show below were obtained by brilliant PhD students and post-doctoral fellows I had the pleasure to work with: Rodrigue Domga Komguem (co-advised with Prof. Maurice Tchuente and Prof. Fabrice Valois), Angelo Furno (co-advised with Dr. Marco Fiore), Panagiota Katsikouli (co-advised with Dr. Marco Fiore), Diala Naboulsi (co-advised with Dr. Marco

[1]Despite submitting five project proposals and actively seeking industrial collaborations on the subject, I did not obtain any funding on vehicular related topics.

[2]Low-power wide area networks, such as Sigfox and Lorawan, are good examples of downlink-limited networks, while visible light communication solutions generally have a limited uplink.

Fiore), Jad Oueis (co-advised with Prof. Fabrice Valois), and Patrice Raveneau (co-advised with Prof. Hervé Rivano). While this document tells the story of all of them just as much as mine, they will excuse me for relegating them to secondary characters for the next four chapters.

I will begin with a chapter entitled *Not At Home* (Chapter 2), where I will discuss my work on collecting network data. I will describe the difficulties of collecting data as a network operator, and the difficulties of working with mobile operators as a scientist. I will emphasize, not too harshly I hope, the way these tasks are complicated by the increased sensitivity to user privacy issues. The chapter will also address actual problems related to data collection, either by crowd-sensing using mobile devices, or by deploying our own wireless sensors network.

The chapter *Inside Information* (Chapter 3) summarizes my work on mobile network data analysis. By taking the example of large datasets of mobile traffic coming from operators in France, Italy, Senegal and Ivory Coast, I will detail the tools we borrowed from the fields of data mining and machine learning to analyze this data. I will mainly focus on three major tasks: classification of the network behavior, prediction of the network state, and detection of anomalies in the network. As network traffic presents both spatial and temporal variations, our solutions can be applied on both these dimensions. The obtained results can be interesting not only from a networking perspective, but also from a social one: temporal classification retrieves daily and weekly human activity patterns, while spatial classification is proven to be an excellent proxy for detecting land use in urban environments.

In *The Return Journey* (Chapter 4), I will show how classification, prediction and anomaly detection of network traffic can be included in the design of networking solutions. I will provide three examples in this sense. First, I will show how classifying network traffic can help with the problem of assigning mobile edge computing (MEC) facilities, such as virtualized application servers, to the base stations of a mobile network. Second, I will explain that predicting network traffic is essential to properly handle user mobile in future mobile networks based on a cloud radio access network (CRAN) architecture. Third, I will demonstrate that even basic functions, such as user association in a mobile network, can benefit from traffic awareness, especially as mobile architectures evolve towards virtualization.

Finally, the chapter named *The Last Stage* (Chapter 5), builds on the lessons learned throughout this quest and discusses the major challenges I see in the

field for the coming years. This chapter is also where I discuss some longer journeys that I would like to undertake in the future.

Chapter 2

Not At Home

Collecting Network Data

> Bilbo was astonished. The only path was marked with white stones, some of which were small, and others were half covered with moss or heather. Altogether it was a very slow business following the track.
>
> *The Hobbit, or, There and Back Again*
> J.R.R. Tolkien (1937)

The underlying idea of this document, which will appear under various forms in this manuscript, is that networking solutions should take a more data-driven approach. Of course, most networking mechanisms have always been adaptive, based on some kind of direct or indirect feedback, and QoS frameworks have been around for decades. But, without trying to minimize the quality of these solutions, in an operational context this adaptation remained at the level of tweaking a transmission window or marking some packets for a specific treatment. The recent advancements in virtualisation increased the number of degrees of freedom available to network functions. This allows for a finer grained control of networks, with entire functions that can be activated, modified or deactivated on a per-packet basis. However, enabling this type of control requires a better understanding of the network traffic and its relationship with networking metrics.

Even before joining INSA Lyon, I started working on a large dataset of vehicular mobility[1], in order to conceive new networking solutions for vehicular networks. Having spent my entire PhD thesis showing the limits of an IEEE

[1] http://kolntrace.project.citi-lab.fr/

802.11-based solution in vehicular environments, the analysis of this dataset was a career-changing experience for me. While IEEE 802.11 solutions were indeed failing under high channel load, the data was showing that these situations only appeared for a few tens of minutes per day and only in very limited regions of an urban area[2] [56]. This convinced me that real data should come early in the research process, not only in the evaluation phase, but even in defining the research questions we want to answer.

This was already my research philosophy when I joined the Inria Urbanet team to work on wireless networks in urban environments, and I was very keen in following it in all my research projects. However, I quickly learned that good data, describing the users behaviour or their surrounding environment, is not easy to come around. Despite numerous open data initiatives, the quality of the data shared through these means in usually very low[3] [14]. Moreover, when data is associated to users, it quickly raises privacy concerns [73] and it is rarely open. For these reasons, gathering useful data became an integral part of my research, and this chapter summarizes three contributions in this context. First, I will discuss my collaboration with Orange Labs, which was the main source of data for my research during this period. This collaboration started within the context of the ANR ABCD project (2013-2017), it pursued with the PIA ADAGE project (2016-2018) and it continues today with the ANR CANCAN project, which started in January 2019. A second point discussed in this chapter is the actual collection of data, using different types of devices. Two examples will be provided in this sense: the development of a data collection application for Android devices, and an experimental campaign using a wireless sensor network for intersection monitoring. The former example was the result of the PrivaMov project (2013-2016) funded by the Labex IMU, while the latter is an ongoing collaboration with the University of Yaoundé I in Cameroon.

2.1 Operator Collected Data

There is no doubt that personal mobile communication technologies are among the most successful innovations of our lifetime. An increasing number of people

[2]Sadly, the prevalent view in the automotive networking community is that we need to find a perfect technology before imposing it to all cars. There will soon be 20 years and 800.000 car crash victims in Europe since we started looking for this perfect solution.

[3]Even mature projects, such as OpenStreetMap, have important shortcomings, e.g. visually imperceptible discontinuities in the road network, which block mobility when they are used for vehicular network simulation.

completely rely on mobile devices not only for work, but also for their personal life and entertainment. In turn, the huge popularity of mobile services has lead to an explosion of mobile traffic. The numbers are always impressing [19]: 66% of the world population was a mobile user in 2017, global mobile data traffic grew 17-fold between 2012 and 2017, and a 7-fold growth is expected between 2017 and 2022.

An indirect consequence of the success of this technology is that mobile subscribers represent today a vast fraction of the population, with 5 billion mobile users in 2017 [19]. Also, mobile devices are continuously interacting with the network infrastructure, and the associated geo-referenced events can be easily logged by the operators, for different purposes, including billing and resource management. Combining the two elements above leads to the implicit possibility of monitoring (some would say surveying) a large percentage of the whole population with minimal cost: no other technology provides today an equivalent coverage. While raising major privacy concerns, which I will address later, this rich source of knowledge represents a clear opportunity to many research communities, allowing scaling up studies across disciplines such as physics [60], sociology [70], epidemiology [35], transportation [33], and networking [30].

Practically, a cellular network is composed of two main parts: a Radio Access Network (RAN), which provides wireless access to the individual devices, and a Core Network (CN), which manages all operations needed to transfer voice and data among different portions of the RAN, as well as to and from external networks, including the Internet. A simplified view of the cellular architecture is depicted in Fig. 2.1.

The RAN is composed of base stations (BS), each in charge of one or multiple cell sectors that jointly cover the geographical surface the network serves. End devices connect to the base station overseeing the cell section they are currently located in. Mobile devices may trespass the cell sector boundaries while exchanging data with the RAN, which generates a handover (HO) event to the new serving base station. Moreover, cell sectors are clustered into Location Areas (LA)[4] that represent the spatial granularity at which the device position is known at all times by the cellular network, and it is thus used for paging. As a consequence, devices moving to a different LA are required to inform the

[4]The notion of Location Area, introduced originally in 2G networks, evolved with the development of new generations of mobile networks. Similar concepts, such as Routing and Tracking Area are described in 3G and 4G systems. However, in this manuscript, I use Location Area as a generic term, denoting all these different technical definitions.

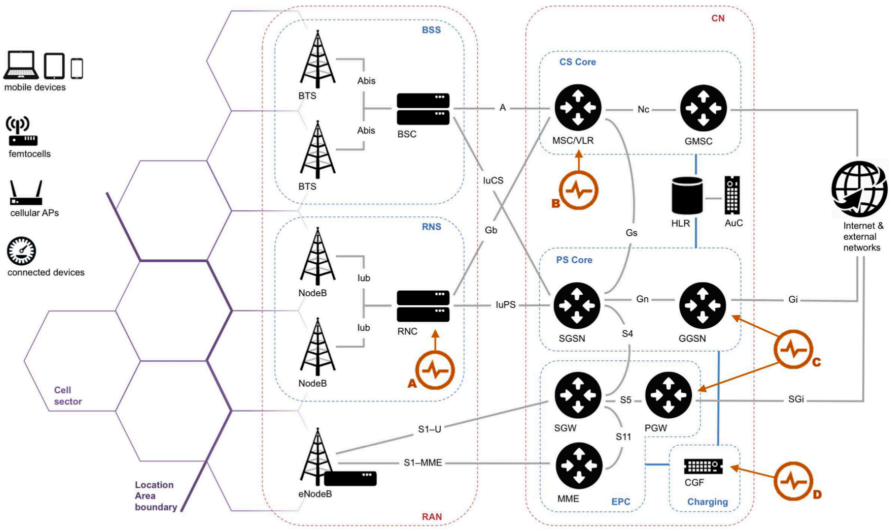


Figure 2.1: Simplified architecture of the cellular network encompassing different 2G, 3G and LTE technologies, and positions of probes for passive monitoring.

network via a location update (LU) event, even if they do not have any ongoing communication at that time.

From a more technical perspective, base stations are referred to as Base Station Subsystem (BSS) and Radio Network Subsystem (RNS) in 2G (GSM, GPRS, and EDGE) and 3G (UMTS and HSPA) architectures, respectively. In both cases, base stations are composed of separated antennas (Base Transceiver Station, i.e., BTS, or NodeB) and controlling hardware (Base Station Controller, i.e., BSC, or Radio Network Controller, i.e., RNC). In the LTE architecture, the eNodeB gathers all base station functionalities.

At the CN, and considering 2G and 3G architectures, voice and texting services are managed via the Circuit Switched (CS) Core, whereas data (i.e., IP-based) services are handled by the Packet Switched (PS) Core. The main entities of the CS Core are the Mobile Switching Center (MSC) and the Gateway MSC (GMSC), which enable voice/text switching within the mobile network and with networks of different operators, respectively. In the PS Core, Serving Gateway Support Nodes (SGSN) and Gateway GPRS Support Node (GGSN) are the interfaces towards the devices and the Internet, respectively, and take care of packet-switched data transfers. In LTE, new entities are introduced to form the Evolved Packet Core (EPC). These manage the control (Mobility

Management Entity, or MME) and data (Serving Gateway, or SGW) planes, and interface them with other IP-based networks (Packet Data Network Gateway, or PGW).

Finally, a set of logical charging function are implemented in the network for billing and inter-operator accounting procedures. 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.

All this abuse of technical vocabulary is needed to explain the different possible placements of data collection probes in a mobile network, depicted in Fig. 2.1. RNC probes, marked as *A* in Fig. 2.1, can be used to capture signaling events concerning any Radio Resource Control (RRC) operation. This allows to record fine-grained state changes of each device, and thus to detect device network attach and detach operations, start and conclusion of sessions, HO and LU events, related to any call, texting, or data transfer activity. Moreover, it allows collecting key performance indicators (KPI) on data transmission, such as the uplink and downlink throughput experienced by the device.

MSC probes, marked as *B* in Fig. 2.1, are similar to RNC probes, in that they can collect similar statistics. However, as MSCs are located in the CS Core, these probes can only track signaling related to voice and texting (and not to data traffic). Moreover MSCs control multiple base stations and thus events that are managed locally by a BSC or RNC (e.g., intra-base station handovers occurring among cell sectors under control of a same BSC or RNC) are transparent to the probe.

GGSN/PGW probes, marked as *C* in Fig. 2.1, tap at links on the Gn/S5 interface of data gateways on the PS Core or EPC. They inspect messages tunneled in the core network via the user data part of the GPRS Tunneling Protocol (GTP-U); this maps to the IP traffic generated or received by mobile devices. Operators typically have measurement infrastructures already in place to monitor IP level statistics from such GTP-U message exchanges. Collected information include the IP session start and end time, device and user identifiers, traffic volume, type of service (i.e., transport- and application-layer protocols,

class of service – such as web, email, streaming audio/video – and name of the application in some cases). In addition, GGSN/PGW probes can associate location information to the data traffic statistics above. To that end, they monitor the control data part of the GPRS Tunneling Protocol (GTP-C), which carries Packet Data Protocol (PDP) Context messages. These messages are transmitted by the SGSN or MME/SGW to the data gateway to establish, update or tear down IP sessions (i.e., IMSI-to-IP address mappings) of end devices. PDP Context messages include, among other attributes, the cell sector where the mobile device is located when the IP session is started or updated, which can be used to localize the data traffic. In current network configurations, no information concerning voice or texting activities can be collected by GGSN/PGW probes.

Finally, CGF probes, marked as *D* in Fig. 2.1, retrieve data from the CGF. The latter is responsible of providing CDR information to the billing domain of the mobile operator, where fees to be charged to the owners of the end devices are determined. It is precisely CDR that are collected by CGF: these contain start timestamp, duration, and originating cell sector of each voice, texting and data traffic activity of every device.

The probes listed above all have strengths and weaknesses. As a general rule, probes located closer to the end devices (i.e., following the alphabetical order in Fig. 2.1) provide a more detailed view of the mobile traffic, but are more difficult to deploy and often less dependable in terms of uptime. As an example, RNC probes deployed at all RNC sites allow observing all significant events occurring in the network, and thus provide accurate information about which cell sector each device is associated to at all times. This represents the ideal data for any study of user mobility or mobile traffic consumption. However, RNCs are geographically distributed, which forces *i)* the deployment and maintenance of a large number of probes to cover a significant geographical area, and *ii)* an important additional long-haul capacity to transfer all events to a central server.

On the contrary, a small number of GGSN/PGW probes deployed at the few data gateways necessary to cover a whole country allows to monitor mobile traffic much more efficiently. In addition, the information provided by such probes gives a rather detailed description of the IP traffic generated by each device, largely sufficient for studies on mobile traffic consumption. On the downside, these probes only yield very approximated positioning information, updated only at the establishment of the PDP Context by an end device[5], or when the device

[5]This maps to the time at which the device opens a data connection to the network. We remark that, once the connection established, a device may keep it open even if it switches to

moves across different SGSN or 2G/3G/LTE coverage areas. The latter events are quite rare, whereas cell sector changes that trigger HO or even LU events – instead very frequent in cellular networks – are not reported up to GGSN or PGW and thus go unnoticed. As a result, GGSN/PGW probes often have stale views of device locations.

The tradeoff is shifted in the case of CGF probes. On the one hand, the CDRs they collect do not provide any insight on the type of data traffic generated by the devices: the rich information on protocol and service-level operations granted by GGSN/PGW probes is lost at CGF probes, which only observe traffic volumes. On the other hand, however, CDRs are readily available to mobile operators, typically at a single server for the whole network, and contain clean, well formatted information on millions of devices. This makes such kind of mobile traffic source extremely popular in research. In addition, the mobility information yielded by CDRs is more accurate than that provided by GGSN/PGW probes, as CDRs include the starting cell sector of each activity, whether voice, texting or data.

Our collaboration with Orange started around data collected by CGF probes, which Orange made available to the research community in the context of the two editions of the Data for Development (D4D) challenge [10, 22]. This collaboration was pursued in a series of three national research projects (ANR ABCD, PIA ADAGE and ANR CANCAN), where we were able to access both RNC probes and GGSN/SGW probes. During the first two projects, only 2G and 3G probes were available, while 4G probes were under deployment. The advantage of using a combination of RNC and GGSN/SGW probes is that this provides accurate information both in terms of user mobility (from the RNC probes) and service consumption (from GGSN/SGW probes).

2.2 User Privacy Considerations

Independently of where the collection probe is located, mobile traffic data contain information on many aspects of subscribers' life, including their activities, interests, schedules, movement, and preferences. It is precisely the possibility of accessing to such information at unprecedented scales that proves of critical importance for studies in many and varied research fields.

an idle state, and thus does not actually transfer data. The device can then become active again, and generate traffic over the same connection that was never closed. This leads to PDP Contexts that are not updated for hours even if the devices change location.

However, accessing such a rich source of information also raises concerns about potential infringements of the privacy rights of mobile customers: among others, individuals can be identified, their movements can be tracked, and their mobile traffic can be monitored. For several years, the common practice adopted by mobile operators in order to protect the privacy of the individuals they monitor was pseudonymisation, also referred to as depersonification. This straightforward approach consists in removing all personal identifiers (e.g., information that is directly linked to the person's identity, such as name, telephone number and different network identifiers), and replacing them with some pseudorandom identifier; the latter can be a keyed hash of the original personal identifiers, or simply a random number that is uniquely associated to an actual individual. Between 2010 and 2015, pseudonymisation was at the basis of several datasets shared by mobile operators, such as Orange, Telecom Italia or Telefonica.

Unfortunately, pseudonymisation only provides a very mild level of protection. A number of experiments (e.g., [46]), performed in recent times and using large-scale real-world datasets, have repeatedly demonstrated the significant risks associated to pseudonymised mobile phone data. In particular, naive cross-correlation of pseudonymised data with named side information (obtained from, e.g., public-access social network data) leads to re-identification, i.e., disclosure of the identities of users with high probability, making pseudonymisation basically useless.

With such a growth of concerns about risks associated with uncontrolled gathering and mining of user data, regulatory bodies have been working on new legal frameworks dedicated to personal data protection. A leading act in this sense is the General Data Protection Regulation (GDPR) [59], which became effective in May 2018 and applies to all European Union citizens. The GDPR enforces that data controllers shall adopt the best measures for data protection by design and by default. Such measures include pseudonymisation, as it can reduce the risks for the data subjects concerned and help controllers and processors to meet their data-protection obligations. However, the GDPR makes it very clear that pseudonymisation alone is an insufficient privacy measure when it comes to sharing and publishing mobile data. Indeed, the regulation decrees that pseudonymised data has still to be treated as personal data, which must be securely stored and cannot be circulated freely. Instead, the GDPR lays down that a more open publication of data is allowed upon anonymization, a process which ensures that the data cannot be any longer linked to an identified or identifiable natural person or data subject. According to the GDPR, anonymized

data is not personal anymore, hence is not concerned by the privacy-protection rules it defines.

However, the problem with the GDPR and other privacy-related directives is that they do not indicate any precise anonymization technique or privacy preservation model to be adopted during or after data collection. Moreover, the situation we have today is that many different notions of privacy exist, not necessarily a subset of each other, such as k-anonymity [71], l-diversity [53], t-closeness [48], and differential privacy [28], to cite just a well-known few. The different regulations and laws do not even clarify which of these definitions should be considered.

The result of such a broad and vague law on privacy protection was a significant one on our work. Research on mobile data has become an administrative nightmare, with interrogation of the researchers by ethical committees prior to any data collection or access, with regular questions coming from national authorities (such as the CNIL - *Commission Nationale de l'Informatique et des Libertés* in France), and with an increased reluctance of sharing data from all the industrial partners[6].

In our case, data collected by Orange probes can be divided in two categories. The first type of data is cell-based, with network traffic information (number of calls, number of messages, downloaded/uploaded bytes) aggregated per cell. The second category is user-based, where the activity of a user can be followed, resulting in a mobile phone trajectory. Cell-based data is anonymous and it is easy to share by the mobile operator; this is the type of data that we mostly used in our studies. User-based data contains the moments and the places (at the granularity of a cell), where a user has been observed in the network. This is considered personal data, since a mobile phone trajectory might allow to retrieve the identity of a user. The mobile operator does not share this data[7], and only allows us access to it on their premises.

2.3 Smartphone Collected Data

Even if user-based datasets would be made available by mobile operators, their usage in human mobility analysis remains limited, because of their reduced

[6]Meanwhile, the business model of web giants built around exploiting personal user data does not seem to have changed, which should raise questions regarding the actual impact of the GDPR.

[7]It is questionable whether the law actually forbids sharing this data, but the Orange legal department does not approve it.

spatio-temporal granularity. Indeed, the mobile user is localized at the precision of the network cell, with a precision in the order of kilometers. Also, the network probes only log the location of the user in case of an event, which sometimes leads to hours (or even days) without any information.

To gather finer grained data, one approach is to collect them directly from the smartphones of the users. Indeed, the large adoption of mobile devices combined to their embedded localization capabilities opens novel opportunities to provide mobility traces to the research community at large. A number of such traces have already been collected, but most of them (e.g., Cabspotting [63], Geolife [85], or T-Drive [81] datasets to name a few) contain data coming from only one sensor (i.e., the GPS, cellular or WiFi interfaces).

Therefore, together with colleagues from other disciplines (economy, transportation, privacy) interested in human mobility, we decided to develop a multi-sensor mobility collection application for smartphones. The idea behind this data collection application was, on our side, to combine the data provided by multiple sensors to increase the precision of user mobility data or to compensate for the lack of data from one sensor, as we will discuss below. In this sense, the decision was to collect data coming from four sources: WiFi and cellular networks, GPS, and accelerometer. The battery state was also logged, for energy consumption optimization purposes.

The PrivaMov application was based on the Funf open sensing framework [2], and several choices were made in order to preserve user privacy and reduce energy consumption. First of all, the application does not open any network interface that the user shut down (e.g. GPS or WiFi). This proved to be an important choice in reassuring users who were not willing to see their data collected 100% of the time. Second, the application does not forces network scans, but simply logs the data every time the system notifies a state modification (e.g. a location change, a new WiFi scan). This tremendously reduced the energy consumption with respect to periodically triggered scans, with the downsize of having temporally irregular information. Finally, the data was collected locally, on the smartphone, and uploaded to our server only when the user was connected to a WiFi network and the battery level was above 80%. While this denied the possibility of having real-time information, it had a positive impact on energy consumption and it did not abuse in any way of the mobile data plan of our users.

The first PrivaMov data collection campaign took place in the city of Lyon from October 2014 to January 2016, in the context of a project funded by the

Sensor	Number of Records
WiFi	25,655,480
Cellular	8,076,512
GPS	156,041,576
Accelerometer	90,066,831
Battery	7,008,504

Table 2.1: Number of records collected by the PrivaMov application during the Lyon campaign.

Laboratoire d'Excellence sur l'Intelligences des Mondes Urbains. For this, we distributed 100 smartphones (52 Wiko Rainbow and 48 Wiko Cink 5) to students and staff from INSA Lyon, ENS Lyon and Université Claude Bernard Lyon. Volunteers were asked to use the phone as their primary phone and to carry it during their daily activities. All the data statistics reported in this section come from this campaign, but the application was also used to collect data during the ACM Middleware 2014 conference in Bordeaux and for a collaboration with clinical psychologists in Saint Etienne in 2016. The application is currently used in the context of the ANR CoWorkWorlds projects, where we study, together with sociologists and urban planners, the mobility habits of people active in co-working places.

Tab. 2.1 shows the number of records collected by the various sensors in the overall dataset during the Lyon campaign. This allowed us to build mobility traces in three different spaces (WiFi, cellular and GPS), which we enriched with automatically detected point of interest (PoI). A PoI is a meaningful location where the user has marked a significant stop, and to extract them we used the methodology described in [84]. The idea behind this method is to identify restricted areas where users stay more than a specific duration. More precisely, PoIs are extracted using a simple spatiotemporal clustering algorithm parameterized with a maximum PoI diameter and a minimum stay time.

Figure 2.2 shows the Complementary Cumulative Distribution Function (CCDF) of the number of unique cells (denoted as Gsm), WiFi access points and PoIs per user, observed in the PrivaMov dataset. We notice a tail distribution, with a few very mobile users visiting a significant number of places.

Having access to more accurate, smartphone collected data allowed us to better understand some of the pitfalls of operator-collected data. One example in this category is the *ping-pong* effect, where a static user switches cellular base

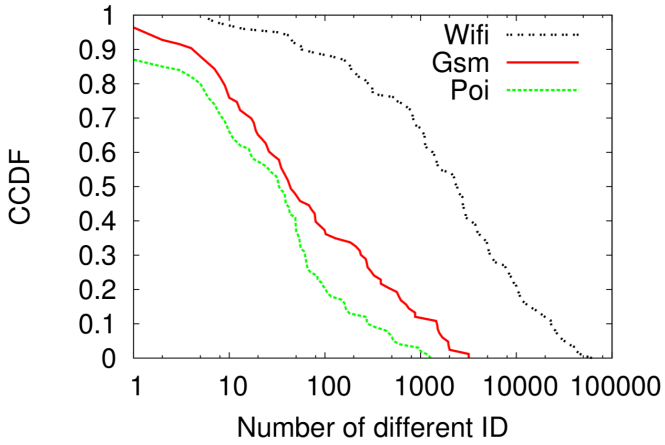


Figure 2.2: CCDF of the number of different cells, WiFi access points, and POI per user in the PrivaMov dataset.

stations. To an unsuspecting observer, this might appear as user mobility, when in fact this is simply an artifact produced by classical mobility management mechanisms in a mobile network. Fig. 2.3 depicts such a scenario, where the user stays within the premises of one building, as shown by the black GPS trace, but the cell tower association changes between three different antennas. By only looking at the red trace in the example of Fig. 2.3, one may reasonably infer that the user is traveling in a circle around a few blocks in the neighborhood. However, that would be an erroneous guess, as the user is actually staying at one specific location. We denote this phenomenon as *oscillations in absence of mobility* and our objective is to remove it from operator collected data, as otherwise this could bias any further mobility-related analysis. Identifying these events is also important from a network point of view, allowing for detailed network planning studies (of course, in that case, we just want to detect these oscillations, not to remove them).

We use data from 11 volunteers where both GPS data collected using the PrivaMov application and operator-data collected by Orange are available. For these users, we divide their GPS trajectories in static and mobile sessions. For ease of reference, we call these ping-pong patterns AXA patterns, where A is the main associated base station and X the (set of) oscillation-produced ones. Usually, X is just one base station that disrupts the sequence of A locations. However, there are also cases (such as the one depicted in Fig. 2.3) where the oscillation involves multiple intermediate base stations. Although AXA patterns may correspond to actual user mobility in periods when the target subject is

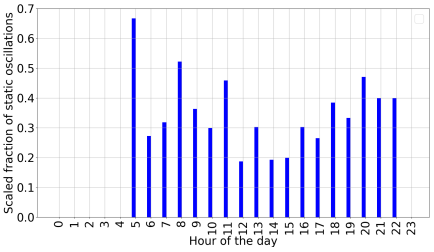
Figure 2.3: Example of an oscillation in absence of mobility. The user follows the cellular trajectory in red, associating to the three base stations. The black line depicts the user non-filtered GPS trajectory.

moving, they certainly are undesired noise when they appear during the stop phases of the movement. Therefore, we are primarily interested in analyzing the presence of oscillations during static sessions.

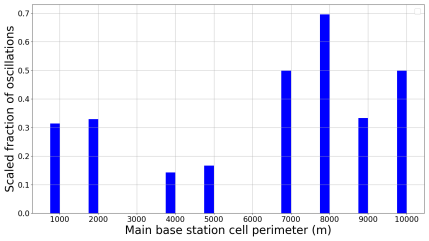
For this, we study the factors that affect the emergence of oscillations in mobile phone location data, during the static sessions detected from the GPS trajectories. In particular, in Fig. 2.4 we consider the impact of the time of the day, the size of the coverage area of the main base station, the density, and the land use of the area where the main base station is located. In all these results, we present the fraction of static sessions that yields oscillation patterns (the results are scaled in the sense that the total number of static sessions extracted for each feature can be different).

From a temporal point of view, Fig. 2.4a shows the impact of the time of the day on the likelihood of observing a static session. The results suggest that oscillations in absence of mobility are more likely to appear in rush hours, especially in the early morning and evening, and are less frequent during work hours. This difference might point to a more relaxed network planning and design in residential areas, where individual users tend to complain less regarding the network quality, than in office areas.

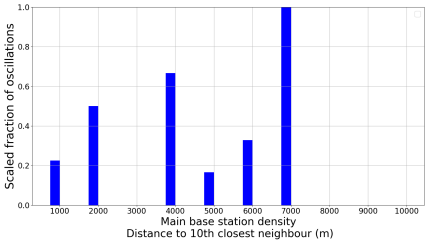
From a spatial point of view, there is one prevalent base station in each static session, appearing more frequently or for the most time. This is the main base station, such as base station A in an AXA pattern. Looking at the properties of this main cell, Fig. 2.4b shows the impact of the size of the cell, given as the perimeter of the corresponding Voronoi area. The probability of oscillation is



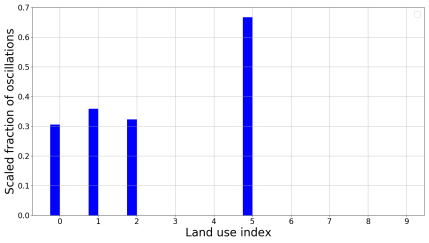
(a) Impact of the hour of the day.



(b) Impact of the size of the main cell.



(c) Impact of the base station density.



(d) Impact of the land use.

Figure 2.4: Impact of different cell characteristics on the fraction of GPS static session presenting a cell oscillation.

the highest in wide cells with perimeters of 7-10 km, where it reaches values up to 70%: the poor signal quality at the (vast) boundaries of these cells is the likely reason. High chances of oscillations, above 30%, are also recorded in small cells, probably due to the vicinity of other base stations that present reasonable network association options for the mobile device.

Fig. 2.4c shows instead the impact of the base station density around the main cell. As a matter of fact, we consider that the distance of the main base station from its 10-th nearest neighbor is an indicator of the density of the radio access infrastructure. In this case, there seems to be some correlation between the occurrence of oscillations and the fact that the main base station is located in an area where the network deployment is sparse, but the trend is not always consistent. Finally, we investigate in Fig. 2.4d whether the land use of the area where the main base station is located has any impact on the detection of oscillations. For this, we use 10 different land use categories: 0 - residential, 1 - mixed office and residential, 2 - office, 3 - central shopping areas, 4 - highways, 5 - large roads, 6 - small roads, 7 - train stations, 8 - leisure areas, 9 - periphery

shopping areas. Interestingly, we observe a high chance of oscillations, above 70%, in the case of large roadways (land use 5), while other categories seem to have the opposite effect, with almost no oscillations. While the statistical significance of these results remains questionable, as very few base stations are found in areas with a land use in categories 4-9, we point out that this also correlates to the temporal results discussed above, where more oscillations are found at times of the day usually dedicated to commuting.

Overall, the previous analysis unveils that no clear-cut rules exist in the appearance of oscillations in absence of mobility with respect to the hour of the day, or to some spatial features of the involved base stations. We found, nonetheless, certain trends as well as the probabilities at which detected patterns can belong to the static or mobile status of the tracked user, which we employ in the design of an oscillation filtering technique. For this, we take a classical approach, where we use 80% of our traces to classify *AXA* patterns as either static or mobile, and we use this classification to filter the remaining 20% of the traces. For the classification phase, we use two techniques, one based on probabilistic classification and the other using a machine learning approach.

In the case of probabilistic classification, we simply use the four features discussed above, namely the hour of the day, the perimeter of the main cell, its density and its land use, to compute the probability that an *AXA* pattern belongs to a static or to a mobile session, denoted as p_s and p_m respectively. The filtering decision is than taken for every pattern where $p_s > p_m$. In the machine learning approach, we use a decision tree based on the same four features, trained, as explained, on 80% of the data.

For the filtering phase, the principle we employ is the following: whenever an *AXA* pattern is found in the mobile phone location data and our classifier labels it as static, we replace the instances corresponding to the oscillation-produced base station(s) X with the main one A, maintaining though the original timestamps of all records.

We evaluate our approach by counting the number of oscillations in absence of mobility before and after the application of the filtering method. We compare our techniques with the state-of-the-art in filtering mobile phone location data from oscillations and noise, i.e. the recursive look-ahead filter (LAF) [44]. The LAF was configured with speed thresholds at 100 m/s and 200 m/s, which are commonly adopted in the literature in order to filter out aberrant trajectories, and at 1.5 m/s, for a fairer comparison with the settings in our proposed methods. Fig. 2.5 summarizes the results. Our techniques remove almost all oscillations

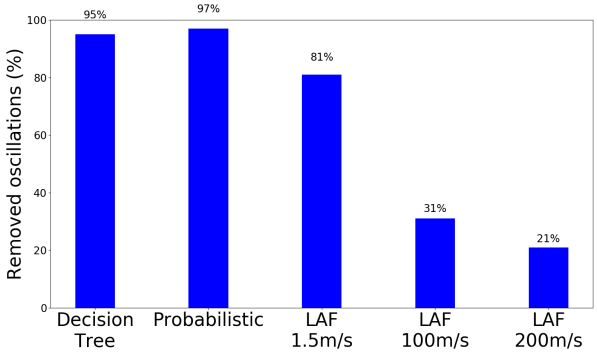


Figure 2.5: Percentage of removed oscillations in absence of mobility.

in absence of mobility, under both probabilistic (97%) and decision tree (95%). The LAF, on the other hand, cannot remove the majority of oscillating patterns (removed oscillations stay at 20-30% of the total), and requires a very low speed threshold in order to achieve a 80% figure.

In order to further assess the quality of the filtering approaches, we design a second experiment. We define an error metric that estimates whether, after the filtering of a mobile phone data trajectory, its distance to the corresponding GPS trajectory has increased or decreased. Namely, we compute the Hausdorff distance between a GPS trajectory and the corresponding mobile phone location data before (denoted by H_b) and after (H_a) the filtering, and compute an error metric as $H_b - H_a$, normalizing the values in the range [-1,1]. Here, a value of -1 indicates a maximally increased distance, hence a negative result implies that the filtering process shifts the mobile phone trajectory away from the ground-truth GPS trajectory. Instead, a value of 1 maps to a maximum reduction of the distance between the filtered mobile phone trajectory and the GPS one.

As we can appreciate in Fig. 2.6, the probability density function (PDF) of this error metric when using the LFA is typically negative, as the LFA does not filter the oscillations, and risks to pick incorrect base stations (i.e., those in X) when it recognizes the AXA patterns. Instead, our techniques exhibit better performance. In particular, with the probabilistic filtering technique, the error metric values become largely positive, indicating that the quality of the filtered trajectory is improved significantly.

The filtering of the ping-pong effect is just one example of the utility of smartphone collected data. This data is not only a complement for operator-

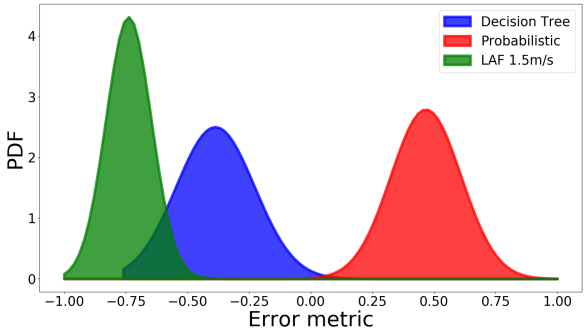


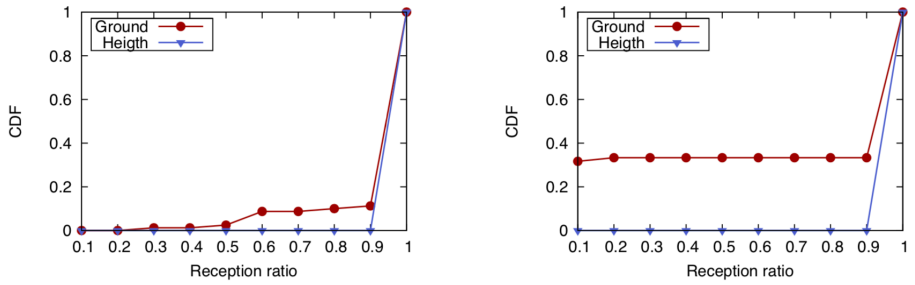
Figure 2.6: PDF of the error metric between the filtered mobile phone trajectories and the ground-truth GPS ones.

collected data, but also an important source of information regarding user mobility and accessed services. But finding users willing to have their data collected is not an easy task, which underlines the interest of collaborations with social scientists, who have already an important experience in recruiting subjects for their studies.

2.4 Data Collection in Sensor Networks

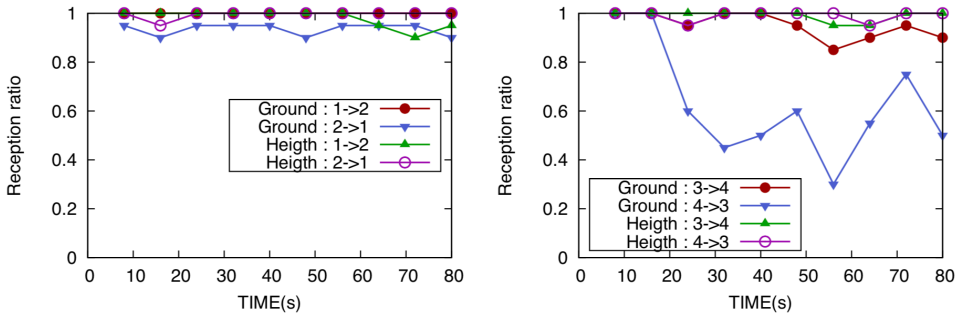
All the previous examples regarding data collection were targeting cellular networks, where an operator is behind a heavy deployment of network infrastructure. However, collecting data is important not only in these scenarios, but also in situations where a lightweight network is deployed, such as in wireless sensor networks.

This was the case in our collaboration with Université de Yaoundé 1, where we conceived a wireless sensor network architecture for monitoring vehicular traffic at intersections. Improving the transportation system is a priority in developing countries like Cameroon, and our architecture highly reduces the deployment cost of a traffic light controller [25]. However, the proposed architecture requires wireless sensors to be deployed on the road, at ground level. The characteristics of such a deployment were previously unknown, meaning that it was difficult to make any assumption regarding the behavior of wireless links in such a situation. Evaluating any network mechanism was complicated in these conditions; we therefore decided to conduct a data collection campaign using TelosB nodes [69],



(a) CDF PRR : Distance 3M (b) CDF PRR : Distance 6M

Figure 2.7: CDF of PRR for distances 3 m and 6 m between nodes in two different deployments: sensors deployed at ground level and sensors deployed on a support of height 57 cm.



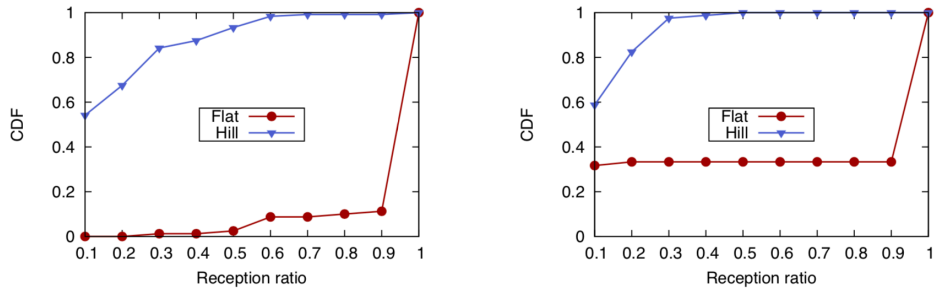
(a) PRR on Link : 1–2 (b) PRR on Link : 3–4

Figure 2.8: Variation of PRR values over time on wireless links 1–2 and 3–4, with the same distance of 3 m, in ground level and at height deployments.

equipped with a 8 MHz microprocessor, a 10 KB RAM, a 1 MB external flash memory and a set of sensors (temperature, humidity and light).

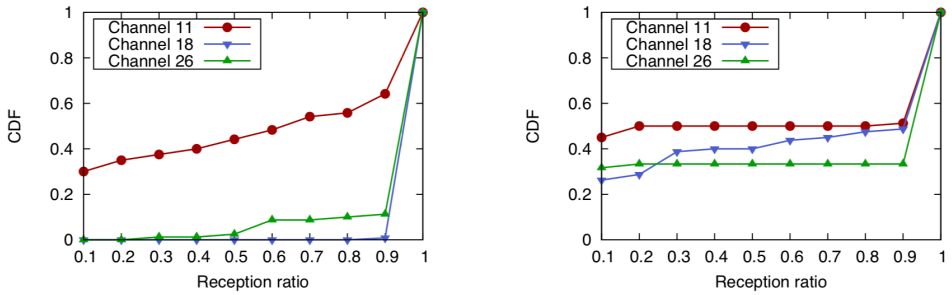
In our experiments, a set of sensors was linearly deployed with a fixed distance between two consecutive sensors. In each experiment, the same communication protocol, avoiding in-network interference, was executed by all nodes. To assess the impact of ground-level deployment on the quality of wireless links, we use two different settings: one with sensor deployed on the road surface, and one in similar conditions, but with sensors placed on a support, tens of centimeters above ground.

In the following, I present a series of figures showing the packet reception ratio (PRR) obtained in different configurations of the network during this collection campaign. First, Fig. 2.7 depicts the PRR obtained on all the links with a distance of 3 m (left) and 6 m (right) between the nodes. At 3 m, the difference



(a) CDF PRR : Distance 3M (b) CDF PRR : Distance 6M

Figure 2.9: CDF of PRR for links with distances 3 m and 6 m between nodes in the case of a deployment on a hill and one on a flat area.



(a) CDF PRR : Distance 3M (b) CDF PRR : Distance 6M

Figure 2.10: CDF of PRR for distances 3 m and 6 m between nodes for different communication channels.

between a deployment at ground level and one at height is minimal, with all the links showing a rather good quality. However, at 6 m, all the links at height have a PRR above 0.9, while more than 30% of the links at ground level have a PRR of 0.

Moreover, the temporal behavior of wireless links is different. Fig. 2.8 shows the evolution of the PRR for two links, the one between nodes 1 and 2, and the one between nodes 3 and 4. We can notice that even the behavior of the same link is different, depending on which node acts as a transmitter, e.g. the communications from node 4 to node 3 has a much more variable quality.

The environment where nodes are deployed also has a significant impact, as shown by Fig. 2.9. In this case, we deployed the nodes in the same configuration, but on two different roads: one flat and one uphill. The PRR in the case of the hill deployment is much lower than in the flat settings.

Finally, we also investigate the impact of the communication channel in Fig. 2.10, where we show results when the nodes are using three different channels in the 2.4 GHz band. Again, the results show significant differences between channels, with channel 11 showing a degraded performance. This is most likely due to increased interference on this channel[8].

All these results show that data collection should take into account multiple factors, not only space, but also topography and external interference. One should be careful not to generalize KPIs collected in particular settings and to diversify as much as possible the considered scenarios.

In our case, the data collected in this campaign was very useful, as it allowed us to define a wireless link model at ground level which we later exploited for two other contributions: a sensor deployment strategy minimizing the overall energy consumption [27] and a self-organization strategy for nodes in a linear network [26].

*****

These projects on data collection helped shaping me as a researcher in many ways. From a methodological point of view, they taught me to be rigorous and to carefully design and prepare data collection campaigns. With a background on simulation and analytical studies, this was a lessons that I learned quickly. Indeed, when you select the KPIs that need to be collected on the Orange network, or when someone installs your data collection application on their phone, there is no turning back and relaunching a simulation in case of a problem. I believe this sense of rigor really became an important element in my research.

The development of the PrivaMov application opened the door to a series of collaborations with colleagues from other fields, especially social and human sciences. Indeed, in these 7 years, I collaborated, in different projects, with researchers in transportation, economy, political science, psychology, sociology, geography, and urban planning[9]. While these collaborations were not necessarily very productive from the publications point of view, I see them as a tremendous chance, which allowed me to broaden my knowledge in all these fields and to discover research problems I never even suspected before. The unexpected

[8]We did not control in any way interference during our experiments, which took place on the open 2.4 GHz band.

[9]In no particular order, I would like to list Pr. Patrick Bonnel from ENTPE, Dr. Olivier Brette from INSA Lyon, Pr. Yveline Lecler from ENS Lyon, Dr. Julie Thomas from Université Jean Monnet, Dr. Nathalie Ortar from ENTPE, Dr. Nicolas Ovtracht from CNRS-LAET and Dr. Patricia Lejoux from ENTPE

challenge when working with people from other fields is a vocabulary one: terms that are well established in our field find different definitions in other fields. Probably the best example in this sense is *mobility*, for which I counted at least four different definitions. After several frustrating meetings, where hours were spent on such misunderstandings, I learned to enter these collaborations with a more pedagogical approach and explain even terms that are obvious in networking.

Finally, from a scientific point of view, data collection brings numerous small problems, that we do not envision when we work on the networking side. For example, in Section 2.2, I briefly explained that we divided a GPS trajectory in static and mobile periods. This might seem simple at first, but at a closer look it raises interesting questions and we spent several weeks designing and validating a proper segmentation methodology. This has certainly changed my vision on so-called simple tasks, especially when these tasks involve using the Apache Hadoop framework[10].

*****

This chapter contains results published in the following articles:

1. Diala Naboulsi, Marco Fiore, Stéphane Ribot, and Razvan Stanica, **Large-scale Mobile Traffic Analysis: A Survey**, IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 124-161, January 2016.

2. Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D'Alu, Vincent Primault, Patrice Raveneau, Hervé Rivano, and Razvan Stanica, **Priva'Mov: Analysing Human Mobility Through Multi-Sensor Datasets**, 5th International Conference on the Analysis of Mobile Phone Datasets (NetMob 2017), Milan, April 2017.

3. Panagiota Katsikouli, Marco Fiore, Angelo Furno, and Razvan Stanica, **Characterizing and Removing Oscillations in Mobile Phone Location Data**, IEEE 20th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2019), Washington DC, June 2019.

[10]Working with this framework was most certainly the worst technical experience in my career. Sadly, it is the framework used by my collaborators at Orange Labs.

4. Rodrigue Domga Komguem, Razvan Stanica, Maurice Tchuente, and Fabrice Valois, **Ground Level Deployment of Wireless Sensor Networks: Experiments, Evaluation and Engineering Insight**, MDPI Sensors. vol. 19, no. 13, pp. 1-25, July 2019.

Chapter 3

Inside Information

Mobile Network Data Analytics

> There were many paths that led up into those mountains, and many passes over them. But most of the paths were cheats and deceptions and led nowhere or to bad ends.
>
> *The Hobbit, or, There and Back Again*
> J.R.R. Tolkien (1937)

In the pre-GDPR era, between 2011 and 2015, more and more network (and not only) data was becoming available. It was the time of big data challenges, not only those organized by Orange [10, 22], but also by other operators, such as the Telecom Italia Big Data Challenge [8]. It was the time when AirSage[1] and an undisclosed French operator were selling individual (pseudonymized) mobile phone data quite cheaply. The term *network science* was beginning to decline, and *data science* to become popular, *machine learning* was still something related to image processing, and *artificial intelligence* was only a thing in robotics. Researchers in sociology, transportation, urban planning, even in physics and in epidemiology, were using all these data to revolutionize their fields, but very few people in the mobile networks research community seemed to take any interest in the subject.

On my side, I already had some experience with large datasets. Actually, my first research experience was an internship, before my PhD thesis, working on a dataset of bike trips. And I was already quite involved in the analysis of

[1] https://www.airsage.com/

the Cologne vehicular trace [76] for my research on vehicular networks. At the CITI laboratory, the Inria UrbaNet team was just created and what we were all dreaming of was to have a large dataset representing realistic human mobility in an urban area, allowing us to simulate lots of wireless network technologies. Using mobile phone data to generate synthetic datasets of human mobility seemed as a logical step, so we started gathering all the data we could get our hands on and analyzing it.

Two things became obvious very quickly: we could not use this data as intended, but its potential was enormous. Indeed, the mobile phone data available at the time, coming from CGF probes, was too sparse to allow the production of a fine-grained microscopic mobility dataset. What we managed to obtain were some mobility flows [57], but this was still very far from the expected results. However, these mobility flows were showing important spatio-temporal variations. The fact that the mobile data demand depends on time and space is not a surprise but, for the first time, we had data allowing us to measure this variation and find correlations and patterns.

This led us to a series of studies analyzing mobile phone data coming from different probes. The common principle of these studies is that they can all be summarized in three steps: *i)* classification of mobile data traffic along the spatial or temporal direction, *ii)* use this classification to predict the mobile network state at a future time, and *iii)* detect an anomaly whenever the prediction is inaccurate. This chapter summarizes these contributions on mobile data analytics, all of which are a result of a collaboration with Orange Labs, in the context of three different national projects (ANR ABCD, PIA ADAGE and ANR CANCAN).

3.1 Temporal Profiling in Mobile Network Data

The fact that user demand in a mobile network varies with time is quite intuitive: most people sleep at night, hence they do not produce any mobile demand, and business-related demand is highly reduced over the week-end. But when exactly does the network switch from one state to another? And what happens if we consider other metrics than the volume, such as the traffic distribution? Is a Tuesday afternoon just a scaled version of a Monday morning?

To answer these questions, in this section, we propose a framework for the classification of mobile network usage profiles. The framework runs on *snapshots* of the mobile demand extracted from any type of available mobile traffic data.

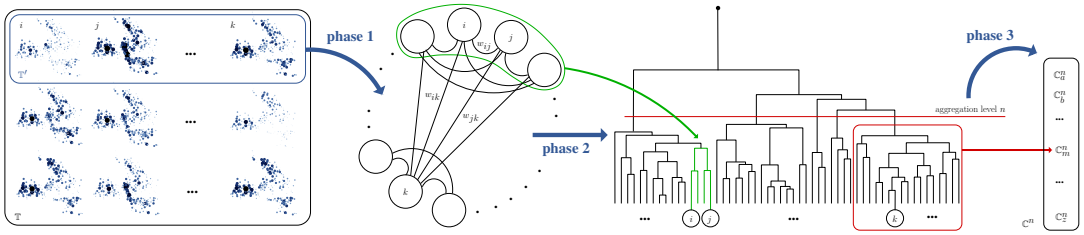


Figure 3.1: Workflow of the framework for the definition of categories of network usage profiles. Phase 1: construction of the snapshot graph from snapshots (portrayed here as geographical plots of the mobile traffic volume) in the training set $\mathbb{T}'$. Phase 2: iterative aggregation of graph vertices into a dendrogram structure. Phase 3: identification of the clustering level n granting the maximum separation between the groups of snapshots. The resulting clusters $\mathbb{C}_m^n \in \mathbb{C}^n$ are mapped to network usage profile categories.

As the name suggests, a snapshot is a representation of the load generated by mobile users on the access network during fixed-size time intervals. We do not impose any constraint on the way snapshots are defined, e.g., they can describe the traffic volume at every second or averaged over longer time intervals, at each base station or aggregated over larger geographical areas, and for one or multiple types of services (voice calls, short text messages, Internet-based applications, etc.). In the following, we will denote as $\mathbb{T}$ the set of snapshots that we aim at analyzing with our framework. Each snapshot will be uniquely identified by the first instant of the time interval it refers to. Similarly, $\mathbb{Z}$ will indicate the set of geographical areas over which traffic volumes are aggregated. At maximum spatial granularity, $\mathbb{Z}$ maps to the set of cell sectors. The choice of $\mathbb{T}$ and $\mathbb{Z}$ may depend on the level of detail of the available mobile traffic data or on the target of the study, yet our framework is general enough to accommodate any definition of such sets.

Once snapshots are defined and extracted from the mobile traffic data, the framework processes them through three phases. These phases aim at defining a limited number of network usage categories by analyzing a training set of snapshots, and their workflow is depicted in Fig. 3.1.

3.1.1 Snapshot graph

In the first step, a subset $\mathbb{T}' \subseteq \mathbb{T}$ of snapshots is selected as the training set over which the categories of network usage profiles are defined. The choice of $\mathbb{T}'$ mainly depends on the available mobile traffic data. For instance, an operator may choose to use snapshots retrieved from the past one-year history to train

the framework, so as to be able to classify the following network usage profiles as they are recorded.

As an example, by using a weekly approach, all the snapshots in $\mathbb{T}'$ referring to Thursdays at 9:00 are merged into a *median Thursday at 9:00* snapshot [2]. Iterating over $\mathbb{T}'$, it is possible to generate a synthetic training set, characterized by the same temporal and spatial granularities as $\mathbb{T}'$ and covering a desired number of days (e.g., one week, by assuming a weekly periodicity of human activity). While we used several options for $\mathbb{T}'$ in our work, the results shown in this document use the median week approach, as it allows for easier to understand visualizations.

Snapshots in $\mathbb{T}'$ are then mapped to the vertices of an undirected weighted graph $G(\mathbb{T}', \mathbb{E})$ that we dub *snapshot graph* (see phase 1 in Fig. 3.1). In the definition above, $\mathbb{E} = \{e_{ij} \mid i, j \in \mathbb{T}',\ i \neq j\}$ is the set of edges e_{ij} between any two snapshots i and j of the training set $\mathbb{T}'$: thus, the snapshot graph is a clique. Each edge e_{ij} is assigned a weight w_{ij}, which is a measure of the similarity between the network usage profiles in snapshots i and j. The way such similarity is measured plays an important role in the framework operation. We propose two different definitions of usage profile similarity that capture complementary facets of mobile traffic dynamics. They are detailed next.

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

The easiest way to compare the traffic volume recorded in two snapshots i and j is to look at the difference of the overall amount of exchanged data, i.e., $\sum_{z\in\mathbb{Z}} v_i^z - \sum_{z\in\mathbb{Z}} v_j^z$, or at measures directly derived from it. In fact, this is a very common approach in the literature (e.g., [67]).

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

$$\mathcal{V}(i,j) = \frac{1}{\sqrt{\sum_{z\in\mathbb{Z}}(v_i^z - v_j^z)^2}}. \tag{3.1}$$

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

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

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

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

$$\mathcal{D}(i,j) = \frac{1}{\sqrt{\sum_{z\in\mathbb{Z}} \left(v_i^z/V_i - v_j^z/V_j \right)^2}} \ , \ V_i = \sum_{z\in\mathbb{Z}} v_i^z \quad \forall i \in \mathbb{T}'. \tag{3.2}$$

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

In our case studies, we use both $\mathcal{V}$ and $\mathcal{D}$ as snapshot similarity measures (i.e. weight w_{ij}in the snapshot graph $G(\mathbb{T}', \mathbb{E})$), as they are complementary in the identification of network usage profiles. In the remaining of this chapter, one snapshot graph is built for each measure, and that the next phases are performed separately on the two graphs (one for $\mathcal{V}$ and one for $\mathcal{D}$).

The snapshot graph is used in the second phase of the workflow as a base for the definition of a set of potential categories of network usage profiles. Here, the goal is to identify all meaningful partitionings of the graph vertices, i.e., snapshots, that display similar mobile traffic conditions. To that end, a hierarchical clustering algorithm iteratively aggregates graph vertices in the snapshot graph into larger clusters, and organizes them into a dendrogram structure (see phase 2 in Fig. 3.1).

We adopt the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [80] – also known as mean or average linkage clustering – as the hierarchical clustering method. UPGMA relies on an agglomerative clustering approach that starts from singleton clusters including one graph vertex each. At every iteration, the algorithm merges the two clusters that share the strongest tie: this means aggregating the groups of snapshots that yield the highest similarity

in terms of network usage profiles. Iterations continue until all vertices are grouped into one cluster.

The dendrogram generated by UPGMA represents a full family of clusterings, as each level in the dendrogram maps to one possible partitioning of snapshots. One must then choose the best clustering, i.e., dendrogram level: the resulting clusters will become our network usage profile categories (see phase 3 in Fig. 3.1).

Many criteria, or stopping rules, have been proposed to automatically detect the best clustering in dendrogram structures: however, there is no clear winner among stopping rules, which may in fact return inconsistent results [54]. In order to achieve a dependable result, we introduce an original criterion, named *top-10 index*, which aggregates the output of seven stopping rules discussed in [54] (the Calinski-Harabasz, Beale, Duda-Hart, C, Hartigan, Krzanowski-Lai and Silhouette indices).

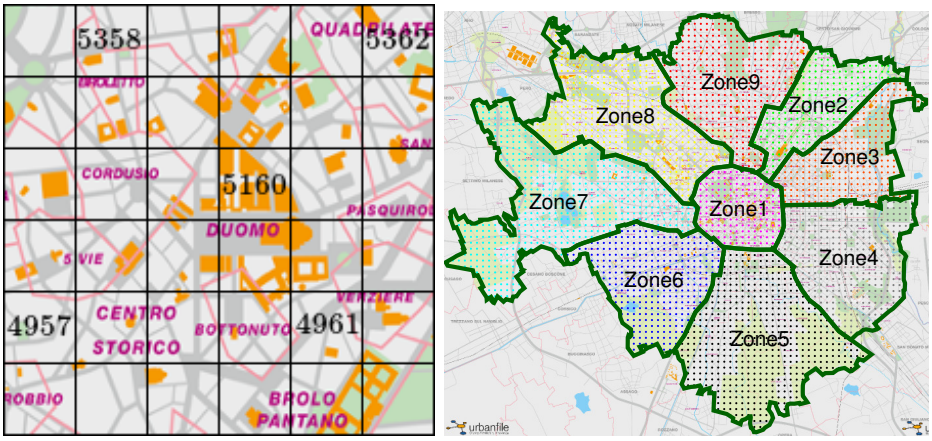
3.1.2 Snapshot classification

Stopping rules allow us to define the aggregation level at which clusters of snapshots show the best trade-off between intra-cluster cohesion and inter-cluster separation. We thus retain the corresponding clustering for our definition of network usage profile categories, as portrayed in Fig. 3.1.

Once the set of categories is identified over snapshots in $\mathbb{T}'$, we can classify the remaining snapshots in $\mathbb{T} \setminus \mathbb{T}'$ accordingly. To that end, we assign each unclassified snapshot to the closest category, where closeness is defined through the same similarity measures introduced in Sec. 3.1.1. We call the resulting category the *actual class* of snapshot i, i.e., $\mathbb{C}_{act}(i)$.

If the training process is performed over the median week, as discussed in Sec. 3.1.1, we can denote the *expected class* $\mathbb{C}_{exp}(i)$ for snapshot i as follows. Let us describe the generic snapshot i by the corresponding day of the week W (Monday to Sunday), time t. Then, we consider the median snapshot x, representing the typical behavior for day of the week W and time t, and select its class from the UPGMA output. The latter is the expected class for snapshot i, $\mathbb{C}_{exp}(i)$. As an example, the expected class for Thursday, 20$^{\text{th}}$ November at 20:00 will be the class of the median Thursday at 20:00 from the median week.

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



(a) (b)

Figure 3.2: a) Division of Milan as a grid zoomed at Milan's Duomo [8]. (b) Administrative zoning of Milan, cells centers.

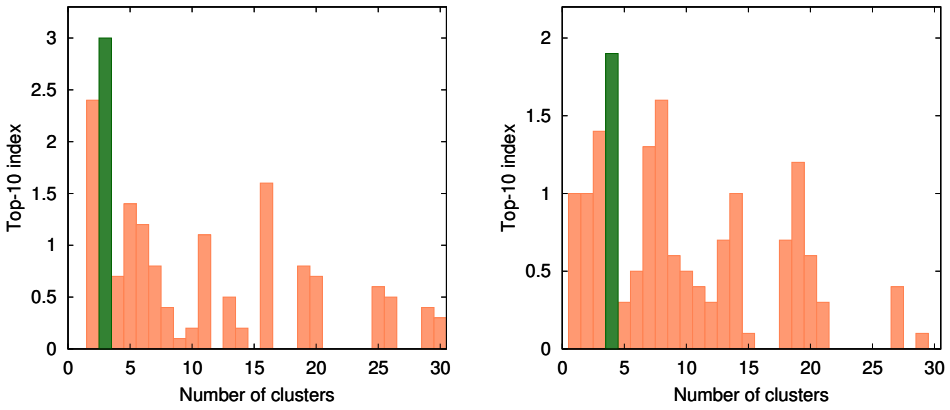
1. if $\mathbb{C}_{act}(i) \neq \mathbb{C}_{exp}(i)$, snapshot i is closer to a different profile than the one of the corresponding median snapshot. In other words, i diverges from the typical behavior in such a way that the associated mobile traffic data resembles those of another profile from the cluster set, according to the selected similarity measure.

2. if $\mathbb{C}_{act}(i) = \mathbb{C}_{exp}(i)$, the distance from the snapshot to the centroid of its cluster is considered as an indication of outlying behavior. Even though the closest cluster matches the expected profile, i can be significantly far from the cluster centroid and, therefore, not well described by the associated class.

In Sec. 3.1.3, we will use two different graphical notations to pinpoint both kinds of outliers.

3.1.3 Results

We evaluated our framework on a dataset provided by Telecom Italia as part of the Big Data Challenge organized in 2014 [8]. This dataset is based on a tessellation of the surface of the city of Milan in cells, see Fig. 3.2a. These cells represent the highest spatial granularity at which mobile traffic measurements are provided by the Italian operator, and they give no information about the deployment of actual base stations in the area. Each square has a 235 m × 235 m size, and the grid is composed of 10,000 squares.



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

Figure 3.3: Top-10 score versus the number of clusters (i.e., the dendrogram aggregation level) for the median week training set.

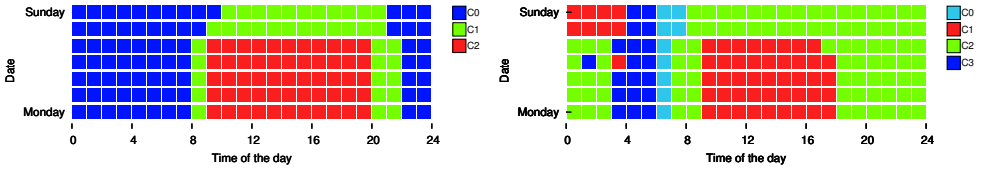
To define the geographical areas $\mathbb{Z}$ for traffic volumes aggregation by our framework, we have used the current administrative subdivision of Milan, made of nine *decentralization zones*. Fig. 3.2b shows these zones. The administrative zones division includes 3,339 cells of the original Milan grid.

The Telecom Italia dataset reports on subscribers' communication activity in terms of received and sent text messages, incoming and outgoing calls and Internet data usage. This data has been extracted from the operator's CDRs and aggregated with respect to each cell of the Milan grid, covering the time period from November 1st, 2013 to January 1st, 2014. Measurements are temporally aggregated in time slots of one-hour. Our final dataset contains 1,488 snapshots.

We exploited our framework to analyze the records related to the different kinds of subscriber activity, i.e., incoming and outgoing calls, incoming and outgoing text messages, and Internet data. Below, I only illustrate results for incoming calls (call-in in the following), which represent an interesting sample of the mobile traffic activity in the urban area of Milan, but I will detail some outliers retrieved by our framework, in relation to the other kinds of activity.

We trained our framework on the median week, computed from the whole 2-month call-in dataset. Fig. 3.3a and Fig. 3.3b report on the top-10 scores versus the number of clusters, computed with the $\mathcal{V}$ and $\mathcal{D}$ similarity measures. They recommend to select *three* and *four* categories, respectively.

The three categories identified by the $\mathcal{V}$ measure, in Fig. 3.4a, are associated with very neat behaviors in relation to call-in volumes. Class C2 groups high-traffic snapshots, mainly related to working and homecoming time, i.e., Monday



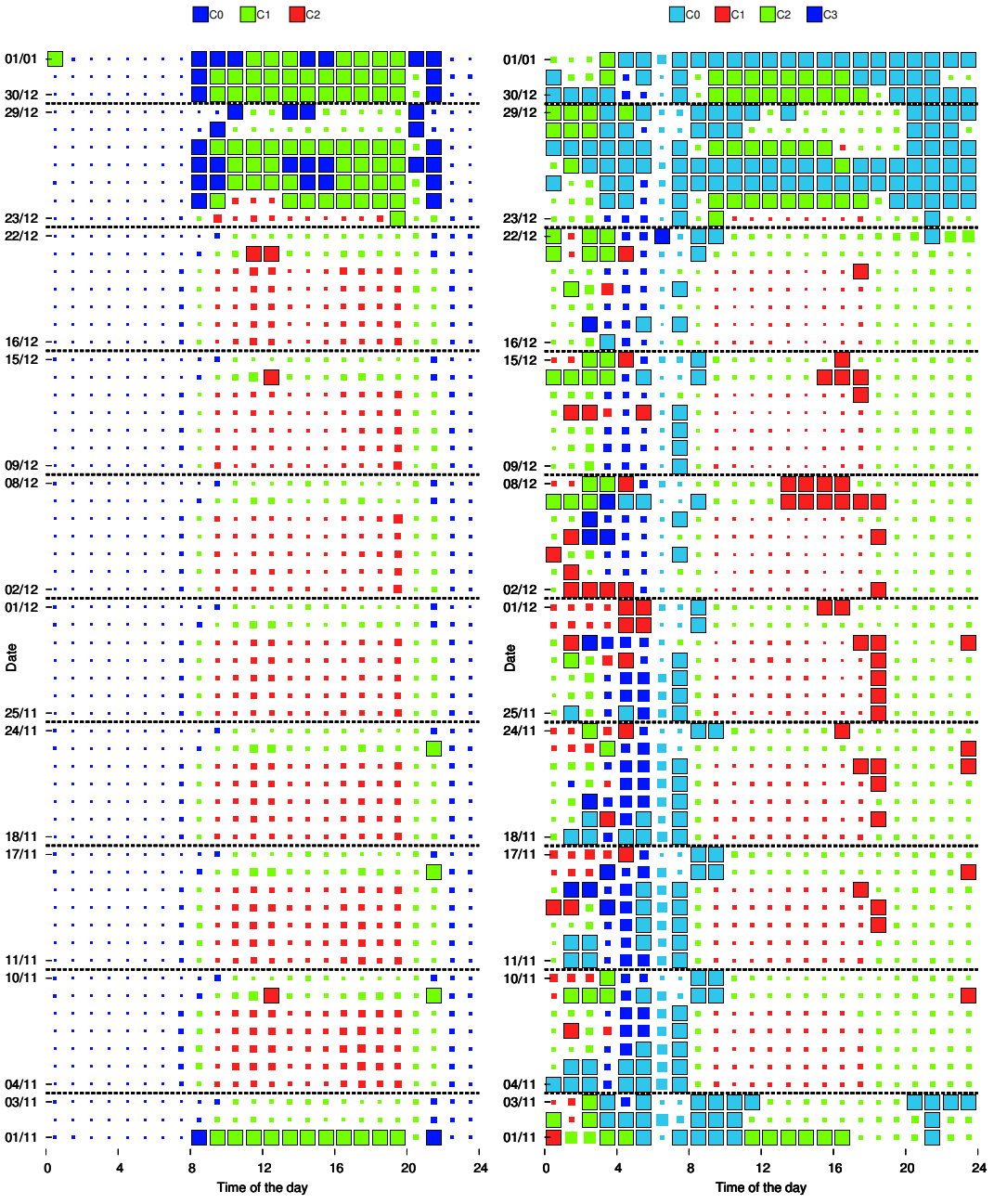
(a) $\mathcal{V}$ - Traffic volume similarity. (b) $\mathcal{D}$ - Traffic distribution similarity.

Figure 3.4: Mobile traffic profile categories defined on the median week. Each square represents one snapshot, and colors map to categories.

to Friday from 9:00 to 20:00. Class C1, contains intermediate-traffic snapshots related to working days, i.e., 8:00 to 9:00 and 20:00 to 22:00, as well as week-end daytimes. Class C0 includes low-activity snapshots related to night and early morning hours, i.e., 0:00 to 8:00 and 22:00 to 24:00.

The $\mathcal{D}$ measure identifies four categories, in Fig. 3.4b, in terms of traffic distribution. Class C1 mainly includes snapshots related to working time, i.e., Monday to Friday from 8:00 to 18:00, and presents the highest concentration of call-in activity in the city center, with a much lower relative traffic in all the other zones. Interestingly, also Saturday and Sunday nights, from 0:00 to 4:00, belong to this class, thus suggesting that night life in Milan is mainly concentrated in downtown. Class C2 is characterized instead by a more even traffic distribution in all the zones of the city. From a time perspective, this class includes night to early morning (18:00 to 3:00 + 1 day), morning (7:00 to 9:00) and most of weekend (9:00 to 0:00) hours. Class C0 is related to early morning hours, i.e., 6:00 to 7:00 during week-days and 6:00 to 8:00 during week-end, with a medium-low traffic concentration in the city center and high demand in Zones 8, 9 and 2. Such zones include important transportation hubs (e.g., metro and train stations) or industrial areas. Finally, class C3 groups deep night hours (3:00 to 5:00), featuring a medium-high load in the city center and a high demand in industrial or high-density residential areas of Zones 2 and 8.

We pursue our analysis by looking at individual snapshots in the 2-month dataset. For this, we iteratively reconduct the median week classification, using only 7 of the 8 weeks and leaving the week of interest out, hence removing the data we want to classify from the training set. The results are very similar to those shown in Fig. 3.4. This allows us to run the classification described in Sec. 3.1.2 on the snapshots of the week left out. We depict in Fig. 3.5 the classification for all the snapshots of the dataset using this methodology, with measures $\mathcal{V}$ and $\mathcal{D}$.



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

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

Figure 3.5: Classification of the 2-month Milan traffic data, call-in activity. The size of the squares denotes the distance to the centroid of the expected class. Solid squares indicate snapshots classified in an unexpected class.

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

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

This allows us to retrieve snapshots with outlying behavior. As explained in Sec. 3.1.2, these can either be snapshots that are classified in an unexpected category (represented by solid squares in Fig. 3.5), or snapshots that are correctly classified, but showing a behavior quite different from the rest of their class (the size of the squares in Fig. 3.5 represents the distance between the snapshot and the centroid of its class). A summary of the most relevant outliers identified in the Milan case study is in Tab. 3.1. The table presents unusual behaviors in the mobile traffic demand, along with their underlying reason, as detected in different types of traffic data (i.e., call-in, call-out, SMS-in, SMS-out and Internet).

Many of these outliers are public holidays, where the users are expected to produce an unusual demand on the cellular network. However, we also manage to detect events on a smaller scale, but with a significant impact on the network traffic, such as football matches or the season opening at the "La Scala" theater.

It is also noteworthy the complementarity of the two metrics, $\mathcal{V}$ and $\mathcal{D}$, which often detect different outliers. The second metric, $\mathcal{D}$, seems more sensitive to specific events. Indeed, this metric focuses on the distribution of mobile traffic in the city, therefore detecting not only moments with an abnormal high (or low) demand, but also moments where users have an unexpected presence in certain areas of the city.

After studying the temporal behavior in different cities and for different services, my belief is that understanding the exact demand fluctuations in each urban region is paramount to a successful deployment and operation of cellular networks. Indeed, the presence of such a strong variability in the way the mobile

Table 3.1: List of outlying snapshots, according to the classification for Milan with $\mathcal{V}$ and $\mathcal{D}$, respectively.

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

network is used underscores the inadequacy of static resource allocation and function deployment, as I will further argue in Chapter 4.

3.2 Spatial Profiling in Mobile Network Data

The mobile network community has always acknowledged that there exist strong relationships between the mobile communication activity and what we refer to as *urban fabrics*, i.e., the combination of infrastructure (e.g., roads, transportation systems, and sports, education, or healthcare facilities) and land use (e.g., residential, industrial, or commercial) that characterizes different zones within a same metropolitan area. Important correlations were found between the mobile demand and the underlying city cartography: notable examples include the spatial diversity of mobile activity within a conurbation [68], the similarity of temporal dynamics of traffic in residential areas [4], or the fact that load peaks undergo geographic shifts between precise urban areas throughout the day [75] and during weekday-to-weekend transitions [78]. Recently, mobile phone data was even leveraged to validate urban planning theories on conditions that promote life in a city [23].

Motivated by these results, we decided to delve deeper in the characterization of the spatial heterogeneity of mobile communication activities. More precisely, we inherited the notion of *mobile traffic signature* to denote the typical activity pattern of the mobile demand at one specific geographic zone [36]. In the following, I will show that such signatures can provide an evident association of prototypical mobile communication dynamics to precise urban fabrics. Moreover, many of these signatures appear to be general in nature, since they emerge in different cities and countries.

3.2.1 Mobile Traffic Signature

Let us consider a generic dataset $\mathcal{A}$, describing the communication activity of a mobile subscriber population during a set of days d. For each day, the mobile demand is stored as the aggregate of the traffic generated by all users in a same area during a given time interval; the size of the area and duration of the interval determine the spatial and temporal granularity of the dataset, respectively. We name *unit area* the spatial aggregation level: the whole geographic region

under consideration, denoted as z is thus divided into unit areas[4] z. The time granularity is instead characterized by the duration of a *time slot*, i.e., the interval during which user activity is aggregated in each unit area. Each day d is thus split into a set of time slots t. Overall, $\mathcal{A} = \{v_z(d,t)\}$, where every element $v_z(d,t)$ describes the total mobile communication activity within each unit area z at time slot t of day d.

The techniques for the construction of a representative set of mobile traffic signatures process the dataset $\mathcal{A}$ through six phases. These phases, detailed below, aim at: *i)* summarizing the mobile traffic activity in each unit areas into a meaningful profile, i.e., the unit area signature (first three phases); *ii)* grouping similar unit area signatures into a limited set of classes, each exhibiting a unique behavior (last three phases).

1. The *signature metric* indicates the nature of subscriber activity to be represented. Examples of metrics are the number or duration of voice calls, the number of SMS, the volume of Internet data traffic, or the kind of mobile services consumed by the users. The metric controls the actual information in each dataset entry $v_z(d,t)$.

2. The *signature support* is the time interval over which the signature is defined. Denoted as a set of days $\boldsymbol{\delta} = \{\delta\}$, the support entails the level of compression of the data into the signature. It can range from a couple of days (implying a high level of compression, since datasets typically span weeks or months) to the entire observation period.

3. The *data denoising* component extracts information deemed to be representative of the typical mobile traffic activity in a unit area, isolating it from the inherent noise in the data. In cases where the signature support is smaller than the observation period, implicit denoising is realized through compression, which increases data robustness by merging multiple $v_z(d,t)$ samples into a single value.

4. The *signature normalization* makes signatures independent from the absolute volume of mobile traffic recorded at a unit area. This allows comparing the mobile communication activity at different unit areas on the sole basis of the mobile demand variations.

[4]The definition of unit area is general, and can accommodate any tessellation of space. Unit areas can map to, e.g., cell sector boundaries, coverage zones of base stations, Voronoi cells, or elements of a grid.

5. The *signature pairwise distance measure* determines the degree of similarity of two signatures.

6. The *signature clustering algorithm* groups together signatures that are alike, leveraging the distance measure above. Ultimately, this last phase returns a set of classes of archetypal signatures, denoted as c. Each class $c \in \mathrm{c}$ maps to a distinct type of human activity.

After a series of test on real data, we propose our own definition of a mobile traffic signature, denoted as Median Week Signature (MWS). This definition is based on the fact that it has been repeatedly shown that there exists a strong weekly periodicity in human occupations [15, 83], which implies that most of the diversity in mobile traffic activity occurs within a one-week period. We thus speculate that a signature describing the typical weekly behavior of the mobile demand at one unit area contains the vast majority of the significant information about the nature of that area. This lets us consider a week-long signature, avoiding dimensionality problems in presence of long time series (which can instead affect [43]), and not discarding any important knowledge (an issue in highly-compressed solutions such as [18]). Our tests also show the median to be a more reliable statistical measure than, e.g., the average or the absolute values, when it comes to assessing the typical activity in mobile traffic. As a matter of fact, the median is much more robust to outliers, which are frequent in mobile traffic due to special events of social, political, sports, or cultural nature [68].

Coming back to the six steps defined above, the metric we use is the sum of voice and text activity volumes, as assumed by all techniques in the literature. The support in MWS is a one week period, representing a median week, i.e., $\boldsymbol{\delta} = \{\text{MON,TUE,WED,THU,FRI,SAT,SUN}\}$. By using the same notation as above, the element associated to time slot t of day $\delta \in \boldsymbol{\delta}$ in the signature of unit area z is

$$s_z(\delta, t) = \mu_{1/2}\left(\{v_z(d, t) \mid d = \delta\}\right), \quad \forall z \in \mathrm{z}, \tag{3.3}$$

where, $\mu_{1/2}(\cdot)$ represents the median of the set within parenthesis.

After extensive tests, we conclude that the best results are obtained when mobile traffic signatures undergo a standard score normalization. To that end, each element obtained in (3.3) is normalized with respect to the mean and standard deviation of all elements referring to the same unit area. Formally, for

a generic element of unit area z

$$\hat{s}_z(\delta, t) = \frac{s_z(\delta, t) - \mu(\mathrm{s}_z)}{\sigma(\mathrm{s}_z)}, \quad \forall z \in \mathrm{z}, \tag{3.4}$$

where $\mu(\mathrm{s}_z)$ and $\sigma(\mathrm{s}_z)$ denote the mean and standard deviation of the set of elements concatenated in the signature s_z.

The MWS is then defined as the concatenation of time-ordered samples:

$$\hat{\mathrm{s}}_z = \Big\|_{\delta} \left(\Big\|_{t} \hat{s}_z(\delta, t) \right), \quad \forall z \in \mathrm{z}. \tag{3.5}$$

In (3.5), $\|$ indicates the time-ordered concatenation of all elements in a set: $\hat{\mathrm{s}}_z$ is thus the concatenation of all elements computed at every time slot during the average working day and the average weekend day.

Finally, our results also show that a distance based on the Pearson correlation coefficient outperforms other approaches. This solution, where the median week is used as a signature support, the standard score for normalization and a Pearson-based distance, is denoted as **MWS-stdscr-pearson** and it is used to obtained the results shown below.

3.2.2 Datasets

We leverage the **MWS-stdscr-pearson** technique to extract meaningful classes of mobile traffic signatures in a substantial set of urban scenarios in Italy and France. Such a study allows characterizing mobile communication dynamics and their intertwining with the urban landscape with high accuracy, across diverse cities and countries.

Our datasets describe the mobile communication activity recorded in four major cities in Italy i.e., Milan, Turin (**Tu-15**), Rome (**Rm-15**) and Trento (**Tn-13**), as well as in six major cities in France, i.e., Paris (**Pa-15**), Lyon (**Ly-15**), Marseille (**Ma-15**), Toulouse (**To-15**), Lille (**Li-15**) and Bordeaux (**Bo-15**). For the specific case of Milan, we consider two separate datasets, **Mi-13** and **Mi-15**, related to two different time periods. This sums up to ten urban scenarios and eleven datasets. Tab. 3.2 labels the datasets and summarizes their main features.

In all case studies, we consider a geographic region of 150 Km^2 around the city center. There, we collect time series of the mobile traffic demand generated by subscribers of major operators, i.e., Orange in France and TIM in Italy.

Table 3.2: Labels and details for the reference mobile traffic datasets.

Label	Source dataset	City	Unit Areas	Period
Mi-13	TIM 2014	Milan	2763 cell grids	Nov/Dec 2013
Tn-13	TIM 2014	Trento	152 cell grids	Nov/Dec 2013
Mi-15	TIM 2015	Milan	434 cell grids	Mar/Apr 2015
Rm-15	TIM 2015	Rome	341 cell grids	Mar/Apr 2015
Tu-15	TIM 2015	Turin	257 cell grids	Mar/Apr 2015
Pa-15	Orange	Paris	1634 base stations	Aug–Nov 2014, Mar 2015
Ly-15	Orange	Lyon	278 base stations	Aug–Nov 2014, Mar 2015
Ma-15	Orange	Marseille	188 base stations	Aug–Nov 2014, Mar 2015
To-15	Orange	Toulouse	220 base stations	Aug–Nov 2014, Mar 2015
Li-15	Orange	Lille	156 base stations	Aug–Nov 2014, Mar 2015
Bo-15	Orange	Bordeaux	158 base stations	Aug–Nov 2014, Mar 2015

The source data provided by Orange consists of CDR describing hourly volumes of voice and text message activity in the whole France, on a per-antenna basis. In French urban scenarios, our unit areas map the coverage zones of the mobile network antennas, which are approximated as the cells of a Voronoi tessellation. Time slots span one hour, as this is the maximum precision granted by the data. The communication activity in the data covers the period from August 12 to November 30, 2014, and from March 3 to March 25 2015, for a total of 132 days.

The TIM datasets are the same we used for the evaluation of the temporal profiling framework, described in Sec. 3.1.3. For the sake of consistency, we formatted the data so that it conforms to that provided by Orange in terms of temporal granularity, i.e., we aggregated traffic into 1-hour time slots.

3.2.3 Overview of signature classes

We use the **MWS-stdscr-pearson** methodology to determine mobile traffic signature classes for the 6,581 unit areas that cover the urban regions in the reference datasets of Tab. 3.2. Overall, a set of 514 classes is identified across all cities.

Fig. 3.6 provides an overview of the signature classes returned by our methodology. The plot shows how classes (rows) are distributed across cities (columns). Colors map to the percentage of city unit areas belonging to a specific class: the darker the color, the more dominant the class within the urban region (see the color range on the right of the plot for the precise value). For the sake of clarity, we limit the plot to the largest 30 classes, i.e., those including at least 10 unit area signatures (see the cardinality of each class on the right listing). These classes account for 75% to 98% of the total surface in each city (see the

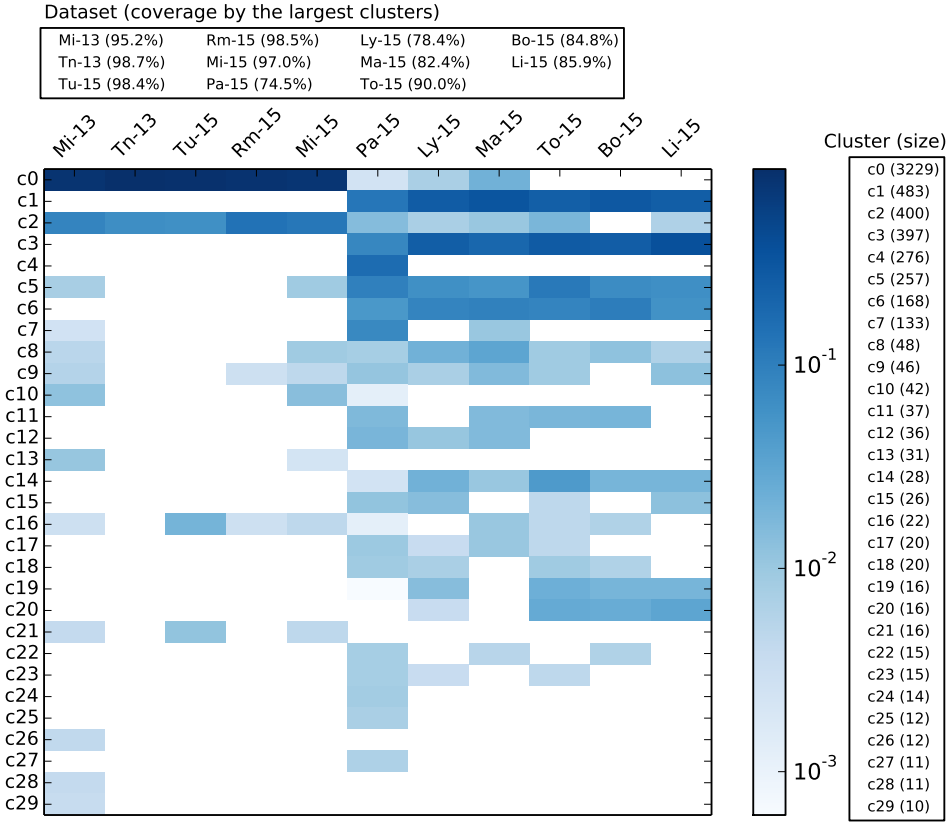
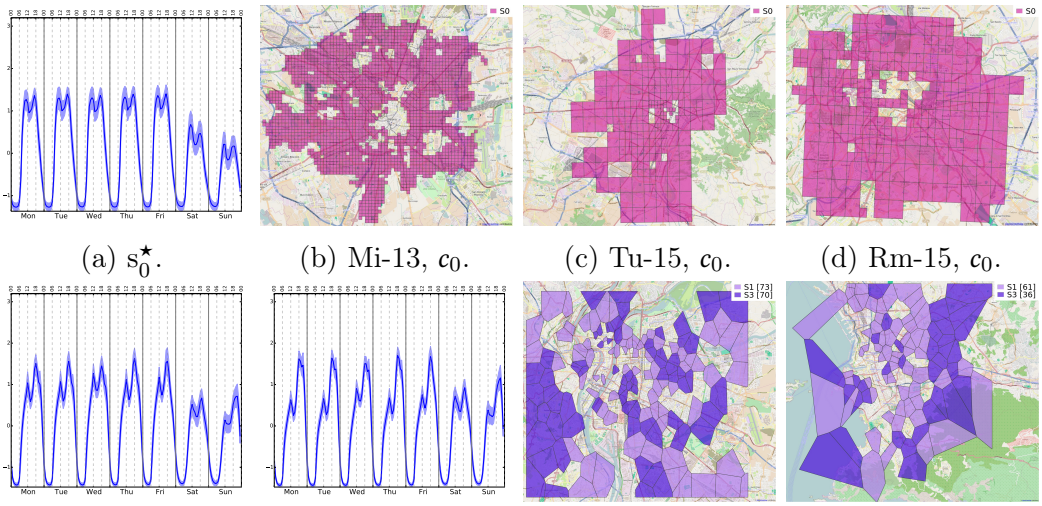


Figure 3.6: Prevalence of signature classes in the reference urban regions.

percentage of city surface covered by the 30 largest classes in the top listing). However, in our discussion of the results, we will also present smaller classes that correspond to peculiar communication dynamics emerging in specific unit areas.

Some preliminary considerations on the signature classes are in order: *i)* class c_0 covers most of the Italian cities, while its presence in France is almost negligible; *ii)* the majority of the analyzed French cities are mainly covered by classes c_1 and c_3, which do not include instead any area of the Italian cities; *iii)* some classes, such as c_2, c_8, c_9 and c_{16}, appear in almost all reference cities, independently of the country; *iv)* other classes, e.g., c_4 and c_{10}, are very city-specific; *v)* Paris displays the highest heterogeneity of classes, and only 74.5% of its surface is covered by the 30 largest classes (the minimum percentage in all scenarios); *vi)* Trento and Rome show the least diversity, as the signatures of their unit areas almost exclusively end in classes c_0 and c_2, with a 98.7% and a 98.5% coverage by the 30 largest classes, respectively.

In the remainder of this section, we will explore the causes behind the classification features outlined above, and more. To that end, we will leverage



(a) $s_0^\star$. (b) Mi-13, c_0. (c) Tu-15, c_0. (d) Rm-15, c_0.

(e) $s_1^\star$. (f) $s_3^\star$. (g) Ly-15, c_1 and c_3. (h) Ma-15, c_1 and c_3.

Figure 3.7: Residential urban fabrics. Characteristic signatures (with standard deviation) and maps of related unit areas in representative city scenarios.

the notion of a *characteristic signature* for each signature class, which can be interpreted as an average of all **MWS-stdscr-pearson** signatures contributing to a specific class c_i.

One remark is that unit areas in French cities appear to be more distributed across signature classes in Fig. 3.6. The consistency of this behavior lets us speculate that the preprocessing enforced on the TIM datasets may have induced important information loss, flattening the diversity of mobile traffic activity. Specifically, the source data for Italian cities has a spatial granularity that can be quite coarse in some cases: Milan grid cells (i.e., unit areas) yield the highest resolution, and the fact that the **Mi-13** and **Mi-15** datasets rank as the most heterogeneous among Italian ones corroborates our conjecture that the spatial discretization introduces some bias in the data. Still, as we will see, this does not prevent the identification of meaningful characteristic signatures in Italian cities as well.

3.2.4 Residential and office urban fabrics

We start our analysis by studying the signature classes that appear the most frequently in the reference urban regions. The characteristic signature $s_0^\star$ of class c_0 is portrayed in Fig. 3.7a. This class characterizes all unit areas of the analyzed Italian cities that do not present any noticeable infrastructure and that do not draw any particular activity of inhabitants. This is outlined, e.g., in Fig. 3.7b,

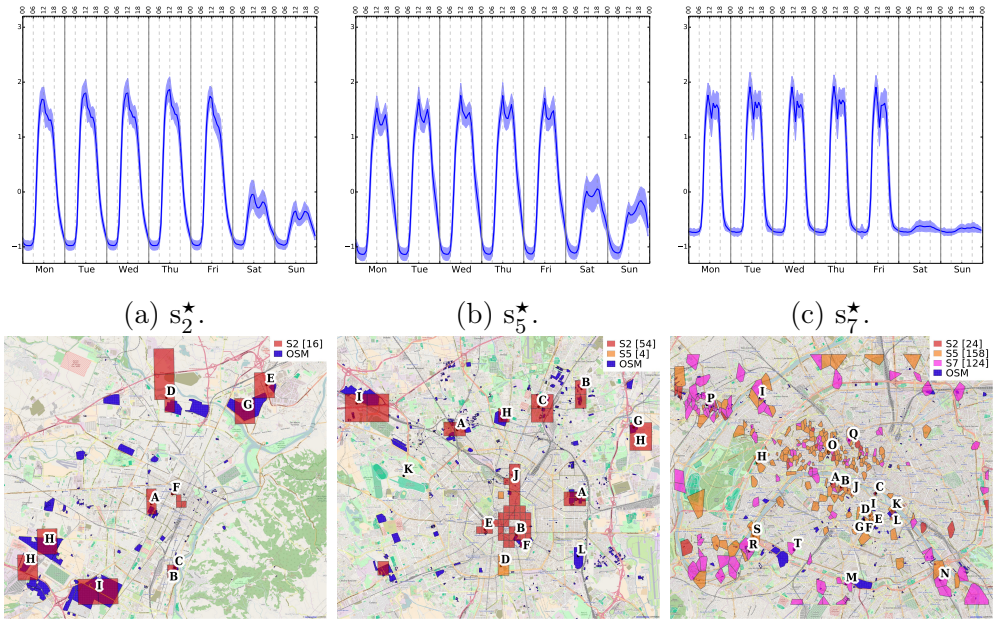
Fig. 3.7c and Fig. 3.7d, which show the extent of unit areas in Milan, Turin and Rome whose signatures are in class c_0. The corresponding regions include suburban and mainly residential areas, and exclude city centers and popular PoIs. Class c_0 can be thus associated to *residential urban fabrics in Italy*, which are denoted by a mixture of private housing and small business activity. It is thus representative of the baseline mobile traffic demand observed within urban regions in Italy.

By looking at the time series in Fig. 3.7a, we remark two comparable traffic peaks, at 11:00 and 17:00, repeating on all working days. The mobile activity in most urban areas in Italy is reduced during weekend, when the morning peak becomes dominant over the afternoon one, which is also shifted towards later hours.

A similar discussion holds in the case of signatures classes c_1 and c_3, this time for French cities. These classes designate *residential urban fabrics in France*, as exemplified by the geographic coverage of the associated unit areas in Lyon and Marseille, shown in Fig. 3.7g and Fig. 3.7h, respectively. The inspection of $\mathrm{s}_1^\star$ and $\mathrm{s}_3^\star$ shapes, in Fig. 3.7e and Fig. 3.7f, respectively, reveals significant similarities in the semantics of the two signatures. Both feature two traffic peaks, the afternoon one standing over the morning one; the activity during weekends is comparable in the two cases, just scaled up in $\mathrm{s}_3^\star$. The main difference between $\mathrm{s}_1^\star$ and $\mathrm{s}_3^\star$ appears thus to be the afternoon-to-morning peak ratio, higher in the latter. We conclude that both c_1 and c_3 are representative of residential and small business areas in France, although c_3 is associated to a higher concentration of residential land use than c_1: indeed, the darker unit areas in Fig. 3.7g and Fig. 3.7h, mapping to c_3, are more present in the urban outskirts and less so in city centers.

It is also interesting to compare c_0, c_1 and c_3. The differences between the baseline profiles of the mobile traffic demand in Italy and France are striking. Activity peaks are uneven and shifted ahead of around one hour in France; their ratio is even reversed during weekends. This diversity is imputable to different routines in the two countries, and entails interesting sociological questions.

On the other hand, unit areas with signatures matching class c_2 are extensively present in both Italy and France, as shown in Fig. 3.6. In order to understand the kind of urban fabrics they pinpoint, we extract layered information on all reference cities from the OpenStreetMap (OSM) database [41] and use it as



(a) $s_2^\star$. (b) $s_5^\star$. (c) $s_7^\star$.

(d) Tu-15, c_2. (e) Mi-15, c_2 and c_5. (f) Pa-15, c_2, c_5 and c_7.

Figure 3.8: Office fabric signatures $s_2^\star$, $s_5^\star$ and $s_7^\star$ and maps of the related unit areas in Italian and French cities, with OpenStreetMap data.

a proxy for land use[5]. When superposing the urban surface covered by unit areas associated with c_2 to OSM data, we remark a good match with locations essentially related to office-hour work activities. Maps of exemplar case studies are provided in Fig. 3.8d, Fig. 3.8e, and Fig. 3.8f for Turin, Milan, and Paris, respectively. A full record of matchings between unit areas in c_2 and office fabrics is instead detailed in Tab. 3.3. Where applicable, capital letters link table entries to maps in Fig. 3.8.

The list is fairly extensive, and we omit a complete discussion for the sake of brevity. As the reader will remark by browsing Tab. 3.3, the signature class appears to highlight office-dense areas, universities, hospitals (especially those linked to research centers or universities), large companies headquarters, administrative centers, and commercial-only areas. We therefore consider c_2 to be representative of *office urban fabrics*, i.e., urban areas interested by socio-economical activities related to development, commercialization and fruition of services and goods, with a typical European working time during week days, 9:00

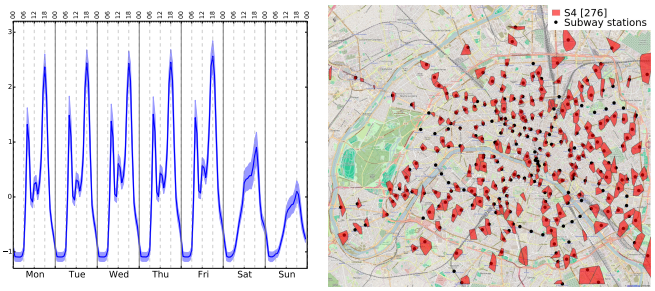
[5]OSM allows tagging geographic areas or even individual buildings so as to denote their primary purpose. The crowd-sourced nature of the information makes it sometimes inaccurate, with tags that can be very generic, associated to multiple activities, or just missing. Thus, we do not treat OSM data as ground truth; rather, it provides hints towards a correct interpretation of the mobile traffic signatures.

Table 3.3: Office PoIs in unit areas of classes c_2, c_5 and c_7.

Class	Dataset	PoIs
c_2	**Tu-15**	Politecnico di Torino (**A**); University of Turin and St. Anne's Hospital (**B**); Le Molinette Hospital (**C**); Telecom Italia Labs (**D**); New Holland Constructions (**E**); Turin Police School, city-center office area (**F**); Iveco Trucks Plant (**G**), Turin Industrial Zone (**H**), Fiat Mirafiori (**I**) plants.
	Mi-15 **Mi-13**	Politecnico di Milano (**A**); University of Milan (**B**); University of Milano-Bicocca (**C**); Catholic University of the Sacred Heart (**E**); Policlinico Hospital (**F**); San Raffaele Hospital (**G**); Mediaset Milano 2 Television Studios and Industrial Pole (**H**); Expo 2015 area (**I**) (Mi-15 only); commercial city-center area (**J**); San Siro Hippodrome Betting pool (Mi-13 only) (**K**).
	Pa-15	Ministry of Defense (**A**); Ministry of Ecology and Development (**B**); Palace of Justice (**C**).
	Rm-15	Ministry of Defense and Intern; Montecitorio Palace, Chamber of Deputies; Policlinico Umberto I; La Sapienza University; Foro Italico University of Rome; Pontifical Universities; Bambino Gesù Children's Hospital; RAI Television.
	Tn-13	Interporto Industrial Zone; Trento Northern Commercial Zone; Spini di Gardolo and Lavis Industrial Zones.
	Ma-15	Aix-Marseille University; Marseille European Hospital.
	Ly-15	University of Lyon I; Parc Technologique Portes des Alpes.
	To-15	Airbus North-West industrial pole; Médipôle Garonne Hospital; Rangueil Hospital; Veolia Peche David plant.
	Li-15	Lille 1 University of Science and Technology; Synergie research-industrial park; Lesquin commercial-industrial park.
c_5	**Mi-15** **Mi-13**	Milano Bicocca Village (Mi-13 only); Milano Bocconi University Campus (**D**).
	Pa-15	Paris I University (**D**); École Normale Superieure (**E**); Curie Institute (**F**); Paris Tech (**G**); Paris Dauphine University (**H**); Paris Sorbonne University (**I**); Paris Descartes University (**J**); Pierre and Marie Curie University (**K**); Polytech (**L**); Ministry of Interior (**O**); Georges Pompidou Hospital (**R**); France Television (**S**).
	Ly-15	Lumière Lyon II and Jean Moulin Lyon III universities; Saint Joseph Saint Luc Hospital center; Le Vinatier Hospital center; Hospital center and University of Lyon Sud; Edouard Herriot Hospital; New Palace of Justice; Cité administrative; Gare de Vaise commercial area; Parilly industrial area; Port Edouard Herriot.
	Ma-15	Timone University Hospital; École centrale de Marseille; Palace of Justice and Courthouse area.
	To-15	Toulouse National Center for Space Studies; Airbus Defence and Space area; Toulouse airport commercial area; Cité de l'espace; Purpan Hospital; Toulouse Sud industrial area; Fondeyre industrial zone; Thales Alenia Space Plant; ON Semiconductors Plant; Freescale Semiconductor Plant; Meteofrance and Aviation Civile centers.
	Li-15	Lille 2 University; Lille 3 Charles-de-Gaulle University; Lille Institute of Political Studies; Pasteur Institute Research Center; Saint-Vincent De Paul Hospital; Regional Hospital University Center; Oscar Lambret, Fontan, Albert Calmette, Roger Salengro and Saint Philbert Hospitals; Les Prés commercial business pole; Marcq-en-Baroeul commercial area.
	Bo-15	Palace of Justice and Courthouse area; French National School for the Judiciary; University School of Management (IAE); Bordeaux University (Campuses Carreire and Talence-Pessac); Pellegrin Hospital; Pessac Commercial area; Bruges Bordeaux Fret and René Ledoux industrial areas;
c_7	**Mi-15** **Mi-13**	Minor industrial area (Mi-13 only).
	Pa-15	Pharmaceutical research laboratories (Sanofi, Pfizer) (**M**); Southern industrial commercial pole (**N**); La Defense business area (**P**); Google France (**Q**); Microsoft France (**R**); Orange France Headquarters (**T**).
	Ma-15	Port Logistic Platform Solaris; Administrative area of Marseille around Place Felix Berret (e.g., Administrative court, Banks and Consulates, etc.).

- 18:00. A confirmation comes from the analysis of the signature $s_2^\star$ associated to c_2, in Fig. 3.8a. The signature is characterized by a fairly constant activity during office hours; more importantly, mobile activities tend to disappear during the weekend, when a very small fraction of offices is open.

It is also interesting to investigate situations where a mismatch is observed between the mobile traffic signature and the OSM data. As an example, an important commercial area is located in Southeastern Milan according to OSM, see L in Fig. 3.8e; however the mobile traffic profile in the area is not that of c_2. Indeed, this is the *Mercato Ortofrutticolo*, a wholesale market where most of the commercial activity is carried out very early in the morning, and that is nearly deserted in the afternoon. Unlike OSM data, mobile traffic signatures neatly detect the quite unique nature of this zone, which is moved apart from standard office areas and into a category per se. Similarly, the Expo 2015 zone in Milan



(a) $s_4^\star$. (b) Pa-15, c_4.

Figure 3.9: Characteristic signature and map of unit areas in c_4. The map includes the locations of subway stations in Paris.

is covered by an office fabric signature in the 2015 dataset, see I in Fig. 3.8e; however it is not in the 2013 dataset, when the area was still under construction.

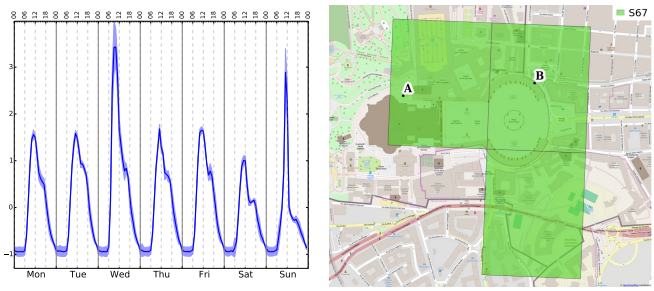
Other popular signatures, mostly located in France, resemble $s_2^\star$. In particular, the $s_5^\star$ signature, in Fig. 3.8b, also presents a quite homogeneous activity with a peak late in the morning, and a high weekday-to-weekend traffic ratio. However, these peculiar features of mobile activity in office areas are blended with those observed for residential fabrics. This suggests that unit areas characterized by c_5 still contain mostly offices, but have a minor presence of residential fabrics. Support comes from Fig. 3.8e, Fig. 3.8f, and Tab. 3.3, as c_5 is mainly located in city centers and other mixed-use areas.

The opposite happens for $s_7^\star$, in Fig. 3.8c, which shows the usual late morning peak, but also a significant reduction of activity at noon during working days and no traffic on Saturday and Sunday. This profile thus denotes pure office fabrics, not contaminated by other land usages. Proofs come from Fig. 3.8f and Tab. 3.3: e.g., in Paris the signature is associated to the Issy-les-Moulineaux area, where headquarters of important companies are located, and La Defense, the major business district of the metropolitan area.

3.2.5 A synthesis of urban fabrics detection

Besides the residential and office urban fabrics, we also detect other major categories, such as transportation hubs (c_4, c_9, c_{10}, c_{36}), or touristic and leisure urban fabrics (c_{14}, c_{16}). For example, Fig. 3.9 shows the signature of category c_4, which maps almost perfectly to the locations of subway stations in Paris.

Out methodology also detects small categories, with a very specific user demand behavior, denoted as *unique urban fabrics*. Many of these are signatures that display dramatic surges in the communication activity that deviate from a



(a) $s_4^\star$. (b) Pa-15, c_4.

Figure 3.10: Characteristic signature and map of unit areas in $c_6 7$, a unique urban fabric covering the area of St. Peter's square in Rome.

regular activity pattern. I can cite here football stadiums (each with its typical schedule of events), or the example shown in Fig. 3.10, the St. Peter's church and square in Rome, with traffic peaks matching exactly the weekly blessing ceremonies of the Pope in that place, which regularly gather a large, diverse audience.

Overall, our signature analysis provides a number of interesting cues that stimulate discussion on the interplay between urban fabrics and mobile traffic dynamics. Below, I summarize the main takeaway messages, separating observations that we find intuitive from insights we deem surprising.

Intuitive findings. Among the expected results, our analysis highlights a clear dichotomy between two prevalent classes of urban fabrics, i.e., residential and office. The former are characterized by a more uniform human presence in time, while the latter – including universities, business centers and company headquarters – are characterized by mobile traffic signatures with high weekday-to-weekend traffic ratios and higher load during typical working hours. Also, one can anticipate that residential areas occupy most of the urban surface, and, thus, that residential mobile traffic signatures are the most common temporal profiles that operators must assume their networks to accommodate.

Another expected result is that touristic and leisure areas are characterized by a relatively high mobile traffic activity during weekends. Equivalent considerations can be made for areas that host periodical events attracting a large number of people, which induce dramatic surges in the communication activity: the likes of arenas, theaters, stadiums or religious places are easily spotted via mobile traffic signatures. Similarly, some nation-specific behaviors could be easily envisaged with minimal knowledge of the local population habits: e.g., activity peak on Saturdays in large French commercial areas and the little or no

human presence on Sundays are easily explained by the fact that commercial centers are typically closed on Sundays in France.

Finally, all the above demonstrates that mobile traffic signatures can detect urban fabrics with a higher level of accuracy than crowd-sourced databases of land use, and possibly pinpoint, at very fine grains, urban zones that yield fairly unique human undertakings. Moreover, mobile traffic analysis allows automatic updating of the urban tissue information, which is not possible with traditional survey-based methods. Both observations were made in previous studies on land use detection based on mobile traffic data, and our analysis reinforces them.

Surprising insights. Less obvious considerations also stem from our results. First, residential mobile traffic in Italy and France shows striking differences, which one would hardly expect from countries that are in geographical and cultural proximity. This diversity lets us surmise that circadian rhythms are intrinsically different in the two countries. We hypothesize that the phenomenon could extend to many major countries in Europe, with important implications in, e.g., the cross-border competition among operators fostered by the new EU regulations on roaming.

Second, such a diversity disappears when moving away from residential city areas. Metropolitan regions that are driven by office, touristic or leisure activities, or that host mass transit infrastructures or main sports facilities have mobile traffic signatures whose traits are consistent in both counties, and across all our reference cities.

Third, ours is the first study to identify mobile traffic activity patterns that clearly denote major transportation hubs. The most distinctive feature is the very high traffic peak in the late afternoon: workers are thus more prone to use mobile services when commuting back home than when travelling to workplaces in the morning. Our analysis also unveils how the mobile demand of medium- and short-range commuters, using subways or urban railways, is semantically different from that of long-range commuters using trains or cars to reach their workplace: the former seem more inclined to mobile communications at all times during their commutes, whereas the latter tend to call and text during return trips mainly.

Fourth, considering ten different cities at once allows us to comment on the diversity observed across them. In the light of the results, our opinion is that the three aspects that drive most of the inter-city differences are *i)* the country *ii)* the size of the metropolitan area, and *iii)* the spatial granularity of the mobile traffic data. We already discussed the dissimilarities in residential mobile traffic

between France and Italy. In addition to this, we remark that Paris, three to ten times larger than all other cities, has unique signatures that tell it apart from the rest. Also, the per-antenna traffic recorded in French cities leads to a higher accuracy (and thus increased heterogeneity) of signatures than that observed in Italian cities. Instead, we do not note major differences between cities of comparable size in a same country.

Fifth, we recall that all our results relate to the mobile traffic activity, which does not necessarily maps to human presence directly. Although the general trends of the two dynamics are probably comparable, we spot noticeable differences in the vast majority of cases. For instance, it is well known that road traffic and railway usage follow a very typical daily pattern, with a high concentrated peak in the morning and a smoother lower peak in the afternoon: our transportation hub signatures are fairly different from this model. Another representative example is that of arenas and stadiums, where major events are characterized by a significant increase of presence, and an even more dramatic surge of mobile traffic demand: individuals attending live shows are keen to digitally share the experience with friends, which exacerbates the network activity peaks recorded in such occasions, with respect to the actual increment of population. Overall, we conclude that mobile traffic signatures are an effective way to detect urban fabrics, but not necessarily human presence.

*****

It is essential to point out at this time that working with mobile network data requires a significant knowledge of the way a mobile network is planned and operated. I strongly believe that a data scientist with no training in wireless networks could not have made any of the contributions we have made in the last years. There are numerous examples I could give in this sense, with hours spent explaining to different collaborators what is a location area, what is a paging message, or why does a base station situated on a hill cover a much larger area than a femtocell placed in a commercial center. With more and more mobile traffic data becoming available, the networking community needs to realize that it is our job to analyze this data and gain insight in the way networks actually perform. Just as we borrowed tools from fields as performance evaluation, operations research, or graph theory, it is time to borrow tools from data mining and machine learning, and apply them knowingly in the networking field.

In my case, the work on mobile data analytics represented a large proportion of my research during the last seven years, with an estimate somewhere close to 50%. I certainly do not consider myself a contributor to the data mining or machine learning fields. But I am clearly a user of these tools, applying them to the analysis of mobile network data. This requires a significant work of surveying the state of the art in the very dynamic data related fields. For example, selecting the best approach in unsupervised learning for a given task requires understanding the theoretical foundations of these different techniques.

From a methodology point of view, working with large datasets proved to be very different from what I was used to do in networking. I was accustomed to define a problem, imagine a solution and evaluate it based on a few well established metrics, like throughput or delay. Integrating data analysis in my work required a very different approach. Of course, when we collect a dataset, we have a certain usage in mind for it. But, faced with a dataset of tens or hundreds of gigabytes, the first step is to understand its properties and check for problems. In one notable example, we realized that a one month data collection campaign conducted by Orange in Lyon was practically useless, as network probes were regularly failing and the dataset presented significant discontinuities. Sometimes, this phase does not uncover problems, but very interesting properties, which produce new research questions, different from those expected at the beginning. The work on urban fabric detection, discussed in this chapter, is one such example.

Working with data also requires a certain interest in visualization. Figures are no longer two-dimensional graphs depicting a simple function. One needs to imagine ways to represent the information in a meaningful way. In our case, working on networking data, this usually involves spatial representations, which explains why geographic information system (GIS) tools, such as QGIS [65], slowly became unavoidable in our projects. I also think that another important element in this case is the capacity to generate and analyze an important number of figures. This allows choosing the best representation of the data, but also understanding important phenomena hidden in the data. Taking once again the example of our study on urban fabrics, thousands of figures were generated in order to observe the mobile traffic signatures and their spatial distribution. In fact, the size of the folder containing the generated figures is higher than the size of the 11 mobile traffic datasets used in the study.

*****

This chapter contains results published in the following articles:

1. Diala Naboulsi, Razvan Stanica and Marco Fiore, **Classifying Call Profiles in Large-scale Mobile Traffic Datasets**, IEEE 33rd Annual International Conference on Computer Communications (INFOCOM 2014), Toronto, April 2014.

2. Angelo Furno, Diala Naboulsi, Razvan Stanica, and Marco Fiore, **Mobile Demand Profiling for Cellular Cognitive Networking**, IEEE Transactions on Mobile Computing, vol. 16, no. 3, pp. 772-786, March 2017.

3. Angelo Furno, Marco Fiore, Razvan Stanica, Cezary Ziemlicki, and Zbigniew Smoreda, **A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas**, IEEE Transactions on Mobile Computing, vol. 16, no. 10, pp. 2682-2696, October 2017.

Chapter 4

The Return Journey

Towards Anticipatory Mobile Networks

> He guessed as well as he could, and crawled along for a good way, till suddenly his hand met what felt like a tiny ring of cold metal lying on the floor of the tunnel. It was a turning point in his career, but he did not know it.
>
> ---
>
> *The Hobbit, or, There and Back Again*
> J.R.R. Tolkien (1937)

Having access to rich datasets allowed us to understand fairly well the way users behave in a mobile network. In fact, our work on mobile data analytics is showing the significant degree to which communication activity patterns are time-varying and location-dependent. As a result, mobile networks must accommodate traffic that shows dramatic fluctuations in space and time, depending on the context of their users.

One might argue that this is why networking protocols and mechanisms are adaptive, that this is why we have back-off mechanisms and congestion control and dynamic address configuration of IP addresses. These are all smart ideas, and actually they are the reason I fell in love with networking in the first place. However, while working with mobile data, I started thinking that maybe we are not doing enough in terms of network flexibility. First of all, there are still entire areas in networking that lack any kind of adaptability. The best example in this sense is the deployment phase, where the planning and dimensioning are static and they can take weeks or even months. Second, seeing that good

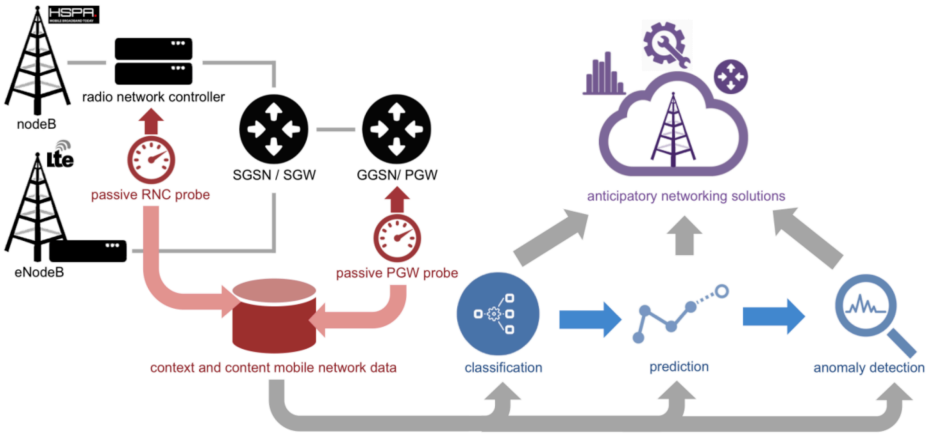


Figure 4.1: Integration of anticipatory networking solutions in the mobile networks architecture.

prediction is possible by monitoring the network, my conviction became that a good networking solution should not simply react to changes, but anticipate them as much as possible.

With the development of virtualization techniques and their integration in networking, this vision of *anticipatory networking* started to become quite popular [13]. The main idea behind this concept is that mobile network analytics, such as those described in Chapter 3, need to be integrated into actual network mechanisms, so as to trigger an informed reconfiguration of the mobile network. The reconfiguration can happen at different levels, and involve network elements, network functions, or network resources. Some recent examples in this sense consist on network paging reduction via user localization [83], RRC protocol parameter tuning [67], or base station switch on/off for energy efficiency purposes [62]. However, while integration in current mobile network architectures is possible in some of the cases above, analytics will best support 5G and beyond-5G systems that are fully compliant with emerging paradigms in network function virtualization (NFV) and SDN [64]. The anticipatory networking paradigm is summarized in Fig. 4.1, building on three major concepts: classification of the network state, prediction of user activity, and real-time anomaly detection.

In this chapter, I will present three contributions we brought in this context. I will start by discussing the problem of infrastructure deployment in mobile networks, arguing that this phase is still stuck in the '80s and definitely in need of some modern solutions. Of course, revisiting the deployment of the radio access and core network brings new challenges, and my focus below is on user association in networks with limited backhaul, such as those where mobile base stations are employed. Our results show that, if one is able to classify the quality of service requirements of mobile users and have a short term prediction of resources on the access and backhaul, the risk of resource outage is significantly reduced with respect to classical strategies.

A second example targets the virtualization of the RAN in the Cloud RAN (CRAN) architecture. By investigating mobility management in this architecture, we observe a new type of handover, previously undocumented, namely the reconfiguration handover. We show that this could have a major impact on user experience if not treated properly. We propose a solution where even a basic prediction algorithm integrated in mobility management mechanisms results in tremendous gains from the user perspective.

The final example discussed in this chapter targets a fully virtualized mobile edge computing architecture, where services partially run on virtual machines close to the user. In order to decide the placement and the migration of these virtual machines, we apply the network profiling framework described in Sec. 3.1. This is the perfect example of how the mobile data analytics we developed can be integrated in the operation of the mobile network the data is collected from.

4.1 User Association in Self-Deployable Networks

Cellular networks have been continuously reshaping communication in our society, through the rapid evolution of standards, products, and use cases. Since the early 1990s, when the first analog generation of wireless cellular technology was replaced by digital communication, cellular networks have not ceased evolving: from voice-centric circuit-switched networks in 2G, to packet-switched 3G, followed by an all-IP 4G, culminating today in the paradigm shifting 5G. However, the careful planning and deployment strategies, implemented by mobile operators in order to provide coverage and data services to users [5], were constants during all these decades. Recent evolutions in terms of communication equipment miniaturization, network virtualization and autonomous vehicles are

challenging this vision. In this context, we imagined a major evolution of the mobile network architectures, through the design of cellular networks that can be rapidly deployed, easily installed, and operated on demand, anywhere, anytime, denoted as *self-deployable* networks.

The development of self-deployable networks emanates from the need of providing users with coverage and data services in a variety of uses cases. A common characteristic of all these scenarios is the requirement for the cellular network to be rapidly operational, without going through a classical planning and deployment phase that can take weeks or even months. First of all, the capacity provided by existing mobile networks is not sufficient when specific events (festivals, sport events, protest demonstrations, etc.) gather important masses of users in some locations, highly increasing the local traffic demand. Second, cellular infrastructure in certain areas might be impacted in the aftermath of natural and man-made disasters; e.g. the tropical storm Harvey, that hit the United States of America in 2017, caused service outage in up to 90% of the cell sites in the affected regions. Finally, professional mobile radio services, such as those used by police and firefighters, need to provide communication services with an unpredictable spatio-temporal demand. On a longer term, future network technologies, from 6G and beyond, can be entirely self-deployable, highly reducing capital expenses.

In this context, self-deployable networks consist of rapidly deployable, easily movable cellular equipment, which takes profit from recent advancements in network function virtualization. Such equipment, denoted hereafter as self-deployable nodes (sdNodeB), is conceived to function both as a self-contained cellular network and within an existing mobile network, embedding both RAN capabilities (e.g., radio signal processing, radio resource management) and CN functions (e.g., session management, routing) [38]. Hence, the sdNodeB can provide network coverage to users in its vicinity without any backhaul connection to an external CN, having access to an entity called Local CN. Similarly, the sdNodeBs can also include some application servers, enabling services (e.g. file sharing) without a connection to an external packet data network (PDN). Our vision is that this new type of equipment can be integrated in aerial and terrestrial vehicles, such as drones, airplanes, robots or autonomous vehicles, or they can even be transported by human users (e.g. firefighters, rescue teams) in a backpack, as demonstrated, for example, by Moradi *et al.* [55].

From an architectural point of view, self-deployable networks distinguish themselves from current cellular network through the fact that a wireless backhaul

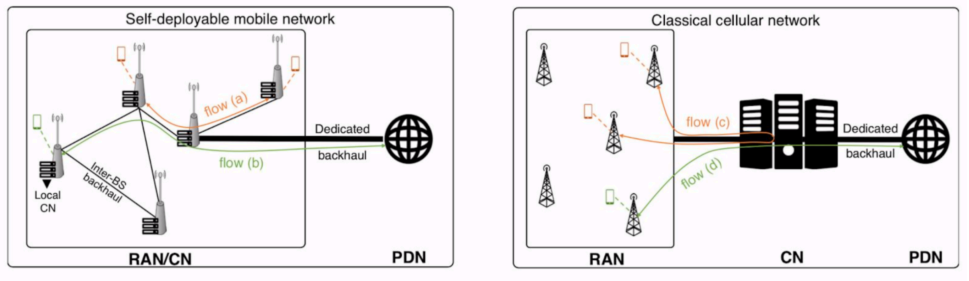


Figure 4.2: Self-deployable and classical cellular network architectures.

needs to be established between the different sdNodeBs, as shown in Fig. 4.2. Since the base stations are mobile, the capacity of this inter-BS backhaul is difficult to dimension and generally limited.

While adding flexibility and democratizing the operation of cellular networks, a self-deployable approach also brings some new problems, such as the optimal placement of the sdNodeBs to cover user demand [52], the placement of CN functions among all the available sdNodeBs [61], or the mobility management of these sdNodeBs [49]. In the following I will address one of the challenges raised by self-deployable architectures, namely the user association problem.

User association, which consists of assigning each network user to a specific BS, is a prevalent problem in cellular networks. Its importance emanates from two prominent challenges faced by mobile operators in general: delivering a consistent quality of experience (QoE) to users, on the one hand, and coping with the scarce available spectrum, on the other hand [31]. In classical cellular networks, the most adopted user association policy is for a user to associate to the BS providing the strongest signal on the downlink (DL) [51]. This simple, yet widely accepted rule, was suitable for the first generations of cellular networks, with a dedicated and over-provisioned backhaul and where the network traffic used to be mostly asymmetric, with the DL traffic significantly surpassing the uplink (UL). However, as we have seen, these properties no longer stand in self-deployable scenarios. Indeed, with a wireless backhaul, critical bandwidth limitations may unfold [24]. When BSs are interconnected, the user traffic is routed on the inter-BS backhaul links that may potentially have a limited bandwidth. Therefore, user association should be adapted to such network architectures, and the inter-BS backhaul state should be taken into account to avoid its saturation.

4.1.1 System Model

To study a self-deployable network scenario, we consider a cellular network based on the integrated access and backhaul architecture, with interconnected sdNodeBs having the ability to act as relay nodes [1]. We consider that each sdNodeB can host, if necessary, (a subset of) CN functions and that there is one sdNodeB with a dedicated backhaul link towards an external PDN, acting as as a point of entry and exit, to and from the network. Let $\mathcal{J}$ be the set of sdNodeBs, and $\mathcal{U}$ the set of fixed user equipments (UEs). The major properties of the network are summarized below.

Backhaul network: The links interconnecting the sdNodeBs form the inter-BS backhaul network, and are responsible for forwarding traffic between the sdNodeBs. When two sdNodeBs are linked, there are two directional links between them, one in each direction. Each backhaul link has a finite bandwidth, limiting the amount of traffic that can be carried by that link (also referred to as link capacity). The study of a specific backhaul technology is not in our scope: we assume a generic out-of-band backhauling operation, such that the backhaul and the access links resources are independent [1]. Without loss of generality, we consider that there is no contention among the backhaul links for resource utilization. We assume that potentially interfering links are operating on distinct channels, allowing interference-free parallel transmissions on the links [6]. Let $\mathcal{L}$ be the set of directional backhaul links. A link $l \in \mathcal{L}$ has a limited capacity, denoted as C_l, in bits/second.

Traffic model: We consider two co-existing categories of traffic flows: intra-network and inter-network. An intra-network flow represents local traffic between two UEs that belong to the same network. Both UEs need to be associated to sdNodeBs within the network, and the association of both UEs is within the scope of our interest. An inter-network flow, on the other hand, is destined to (or originated from) another network. It is a flow between a UE belonging to the network, on one side, and an external UE or an external PDN (e.g. an application server or the Internet), on the other side.

In our formulation, an inter-network flow can be considered as a particular case of an intra-network flow, with only one UE to associate within the network (instead of two UEs). This flow needs to be routed between the sdNodeB selected for association by the UE and another sdNodeB representing the network entry/exit point. In our study, we consider only one sdNodeB acts as a network entry/exit point, with a dedicated backhaul link towards an external PDN.

Everything happening beyond this sdNodeB (i.e. outside the network) is not of particular interest to us: our goal is to determine only the association of the UE(s) belonging to the network under study. We denote by β the percentage of intra-network flow requests out of all flow requests.

We model all flow requests between two parties u and v as bidirectional, meaning they are composed of two directional flows, one in each direction. The throughput requirements can be different on each direction, and are denoted by d_{uv} and d_{vu}, respectively. Flow requests arrive between two random parties according to a Poisson process, with an average flow arrival rate λ_f and a flow duration following an exponential distribution of average μ_f. To study the impact of the increasing UL traffic, we consider that bi-directional flows can be asymmetrical, with α representing the percentage of UL traffic out of the total traffic in the network.

One UE can have several simultaneous flows either intra-network or inter-network. A UE with no ongoing flows is not associated to any sdNodeB. By default, we use a joint DL and UL association, in which a UE is associated to one and only one sdNodeB. However, we also study the impact of a decoupled DL and UL association, as imagined by recent studies [29].

Routing on the backhaul: When two UEs are associated to the same sdNodeB, the data they exchange does not need to be routed on the backhaul. Otherwise, data traffic is routed on the backhaul links between the two implicated sdNodeBs. Since not all sdNodeBs pairs are necessarily directly connected by a link, a routing path between the two end sdNodeBs must be determined. A routing path, denoted $P(i,j)$, consists of a succession of directional backhaul links on which traffic is routed from sdNodeB i to sdNodeB j. Routing paths between two sdNodeBs i and j, i.e., $P(i,j)$ and $P(j,i)$, are not necessarily identical in both directions.

Radio access network: We adopt an orthogonal frequency division multiple access (OFDMA) model, with $\mathcal{K}$ orthogonal channels, divided among the sdNodeBs following a frequency reuse scheme of reuse factor r. The number of channels reserved for each sdNodeB are denoted $\mathcal{K}_j^{DL}$ and $\mathcal{K}_j^{UL}$, on the DL and on the UL, respectively. We suppose that all sdNodeBs are identical in terms of maximum transmit power and antenna gain.

Resource allocation: Typically, for each of its flows, a UE would receive a number of physical resource blocks (PRBs) via a scheduling algorithm. However, to maintain the tractability of the framework, we do not study our system at the granularity of a PRB. Instead, we model resource allocation as an allocation

of a fraction of the available channels to each flow, on the UL and the DL. The number of channels allocated to a UE by a sdNodeB, for each of its flows, depends on its guaranteed throughput requirement and on the data rate the UE observes from the sdNodeB.

4.1.2 Association Policy Overview

Considering the ongoing evolutions of cellular networks, we argue that current association policies, based mainly on the quality of the DL, or the RAN alone, need to be revisited. We propose an association policy comprising two phases: a flow admission control and an association decision. The whole procedure is triggered by the arrival of each flow request.

We assume that re-association is not allowed[1], such that a UE remains associated to the same sdNodeB even when a new flow arrives. In the flow admission control phase, for the already associated UE(s), the network checks if there are enough resources on their sdNodeB(s) to accommodate the new flow. For the UE(s) that are not yet associated, the network checks if there are available sdNodeB(s) with enough resources and available backhaul bandwidth to accept the arriving flow.

For each bi-directional flow request between UEs u and v, of data rate requirements $d_{u,v}$ and $d_{v,u}$, the eventual goal is to find a pair of sdNodeBs (j_u, j_v), such that the following constraints are fulfilled simultaneously: *i)* UE u associates to sdNodeB j_u and it can get a throughput $d_{v,u}$ on the DL and $d_{u,v}$ on the UL; *ii)* UE v associates to sdNodeB j_v and it can get a throughput $d_{u,v}$ on the DL and $d_{v,u}$ on the UL; *iii)* there is a routing path $P(j_u, j_v)$ on the backhaul between j_u and j_v that has enough capacity to carry the flow of data rate $d_{u,v}$ and *iv)* there is a routing path $P(j_v, j_u)$ on the backhaul between j_v and j_u that has enough capacity to carry the flow of data rate $d_{v,u}$. When u and/or v are already associated, j_u and/or j_v are given.

If no pair (j_u, j_v) is found, then the flow is rejected. If only one pair (j_u, j_v) fulfills the constraints above, then UEs u and v are associated to sdNodeBs j_u and j_v, respectively. If more than one pair fulfills the constraints, the best pair is selected with the aim of maximizing the remaining network resources, by accounting for both the RAN and the backhaul.

[1]We tested re-association in our study and the results showed a slight gain with respect to a strategy without re-association.

Several input data are needed to make the admission and/or the association decision: the UEs at both ends of the flow, their required throughput, their channel gains with the sdNodeBs in the network, and the network state represented by the available resources on the DL, the UL, and the backhaul. A part of this information can be provided by the UE in the association request message, while the rest (e.g., the network state, especially the backhaul) need to be predicted. Our goal here is not to demonstrate any prediction mechanism, but to evaluate a backhaul aware association policy. From a practical point of view, one way to implement such a strategy would be to have a centralized control entity that takes these parameters as input, from the UEs and the sdNodeBs, does the necessary computations, and outputs the association decision.

4.1.3 Evaluation of Backhaul-aware Association Policy

We evaluate the performance of our proposed association policy, denoted as DUB (for Downlink-Uplink-Backhaul), with respect to a traditional best SINR association, where users associate to the sdNodeBs from which they get the highest SINR. Furthermore, we evaluate similar policies that follow the same procedure as described in the previous section, but differ in the user association decision process. More precisely, to compare with DUB, we consider a DL-based policy (denoted as D), where only the resources available on the DL are considered, and a RAN-based policy (denoted as DU), where both DL and UL resources are taken into account. The target metric is the flow blocking probability, i.e. the ratio of blocked flows over the total number of flow requests.

We conduct the study in a series of scenarios: we vary the percentage of intra-network flows, the flow arrival rate and the backhaul links capacity. We set $\mathcal{K} = 120$ orthogonal channels, to be divided between the DL and the UL. Channels and power are equally divided among the sdNodeBs with a reuse factor $r = 3$. For the channel reuse scheme, we suppose that the set of interfering sdNodeBs on the DL and the UL are identical.

In the following, we consider the network topology shown in Fig. 4.3, with 6 sdNodeBs in a square of area 1 unit square. We note that the presented results were also validated on several other topologies. We suppose that all backhaul links in $\mathcal{L}$ have the same capacity C_l. We consider 100 UEs, randomly distributed in the area. Flows arrive between random UEs following a Poisson process with an average flow arrival rate λ_f, and a duration following an exponential distribution of average $\mu_f = 180$ s. We fix μ_f, in the following, and vary λ_f. We suppose that

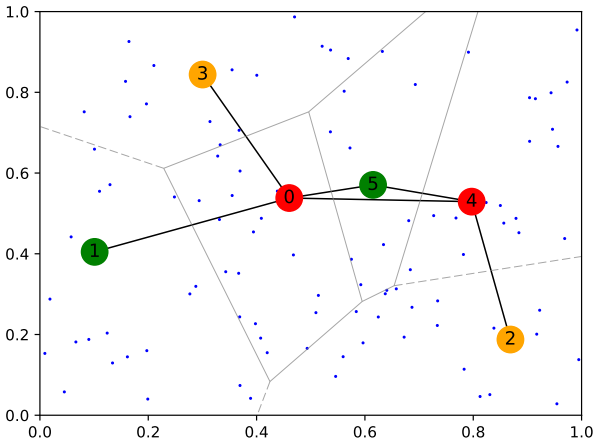


Figure 4.3: Representative network topology used in our tests.

all the flows request a data rate $r_t = 1$ Mb/s, in both directions. Henceforth, all the results are averaged over 30 simulations, with each simulation corresponding to a different user distribution, with 2000 flow requests. The confidence level is set at 95%. All the simulations are conducted using a home-built Python simulator based on SimPy, a process-based discrete-event simulation library.

First, we study the effect of having an increase in local, as expected from device-to-device communication or from applications such as autonomous driving. We do this by varying the parameter β that represents the percentage of intra-network flows out of all flow requests. For now, we set the average flow arrival rate to $\lambda_f = 0.01$ s^{-1}. We test different values of the inter-sdNodeB backhaul links capacities C_l. Results are shown in Fig. 4.4, for $C_l = 1$ Mb/s, and $C_l = 5$ Mb/s. As shown later on, this second value is enough to remove any bottleneck on the wireless backhaul when DUB is used.

In Fig. 4.4, we observe that DUB largely outperforms a best SINR policy, achieving significantly lower flow blocking probabilities, for all values of β and C_l. With best SINR, when the backhaul is limited ($C_l = 1$ Mb/s), the blocking probability is very high, ranging between 30% and 40 % for the different values of β. This confirms that a best SINR policy, oblivious to the backhaul state, is not suitable at all for backhaul-limited networks. The blocking is reduced by a factor of 10 with DUB.

For all values of C_l, the flow blocking probability increases with β. When all flows are inter-network, i.e., $\beta = 0$, the blocking probability is lower, because each flow is only consuming RAN resources on one sdNodeB, as we are only

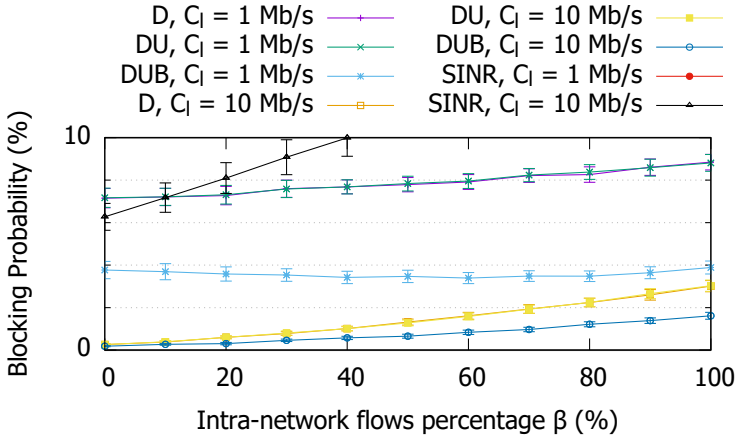


Figure 4.4: Blocking probability function of the percentage of intra-network flows out of all flow requests β, for different inter-BS link capacities C_l. The SINR strategy with C_l= 1 MB/s results in a blocking probability much higher than the 10% threshold represented in the picture.

associating one user per flow. This is not the case for $\beta > 0$, when there are local intra-network flows, since we associate two users per flow, both consuming RAN resources, which increases the load in the network, and consequently the blocking probability.

To summarize, these results underline the poor performance of DL-based association when the backhaul is limited, as well as when a significant proportion of local traffic is present in the network. A careful, more informed association policy, such as DUB, highly alleviates the problem.

Without loss of generality, we focus hereafter in the case β = 100%, which produces the highest blocking probability. We study the effect of the average flow arrival rate λ_f on the association policy performance. We show in Fig. 4.5 the blocking probability of the association policies, function of the flow arrival rate λ_f, for different backhaul link capacities C_l.

We set the maximal accepted flow blocking probability to 10%, as any higher value is superfluous and not tolerable in practice. As shown in Fig. 4.5, the gain achieved by DUB with respect to best SINR is significant, for all values of λ_f, and for the different backhaul links capacities. By comparing the values of the arrival rate λ_f beyond which the blocking probability surpasses 10%, we find that, with a best SINR policy, this value is reached for much smaller arrival rates. Indeed, our results show a gain of 10x in capacity when using DUB for C_l = 1 Mb/s, and a gain of 5x when the backhaul capacity is increased to C_l = 5 Mb/s. This demonstrates, once again, the problems of classical association in

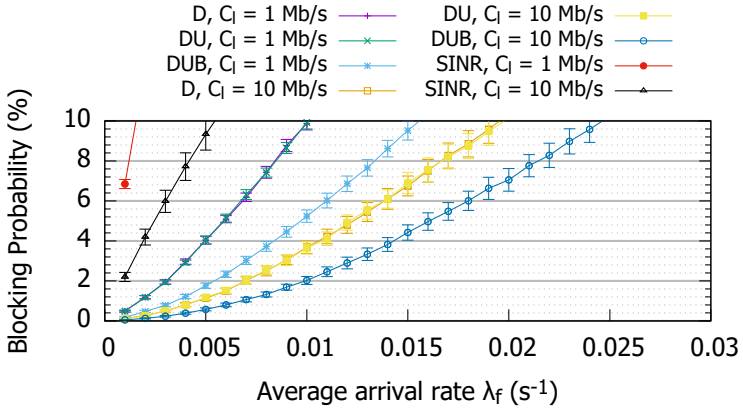


Figure 4.5: Blocking probability function of the average flow arrival rate λ_f, for different inter-sdNodeB link capacities C_l.

limited backhaul scenarios. Indeed, in a best SINR policy, if a UE receives the best SINR from a sdNodeB which does not have enough capacity to accept a new flow, then the flow is dropped. In DUB, the UE is given a wider choice in terms of sdNodeBs, and the association accounts for the remaining resources and the bottlenecks to avoid early saturation of a sdNodeB and/or a backhaul link.

Finally, we set the average arrival rate value to $\lambda_f = 0.01$ s^{-1}, and $\beta = 100\%$. In order to investigate what is causing the flow blocking in each of the association policies, we consider different representative scenarios, that correspond to different backhaul link capacities values C_l, and evaluate the blocking causes. The possible flow blocking causes are: *i)* $\mathbf{BH} = \emptyset$: there is no feasible path on the backhaul; *ii)* $\mathbf{UL} = \emptyset$**:** there is no candidate sdNodeB that has enough resources on the UL; *iii)* $\mathbf{DL} = \emptyset$**:** there is no candidate sdNodeB that has enough resources on the DL; *iv)* $\mathbf{UL} \cap \mathbf{DL} = \emptyset$**:** there is no candidate sdNodeB that has enough resources on both the UL and DL, simultaneously (but UL $\neq \emptyset$ and DL $\neq \emptyset$); *v)* $\mathbf{UL} \cup \mathbf{DL} = \emptyset$**:** there is no candidate sdNodeB that has enough resources on the UL, nor a candidate sdNodeB with enough resources on the DL. Fig. 4.6 shows the blocking probability and the detailed blocking causes for the best SINR and the proposed DUB policy.

We first consider a relatively well-provisioned backhaul, such that all backhaul links have a capacity $C_l = 10$ Mb/s ; $\forall\, l \in \mathcal{L}$ (Fig. 4.6, case (b)). Even in this case, DUB reduces the flow blocking probability by a factor of 6. By investigating the causes that lead to the flow blocking, we notice that, in a best SINR policy, the major blocking cause is the UL. Thus, in this scenario, it is the RAN, and

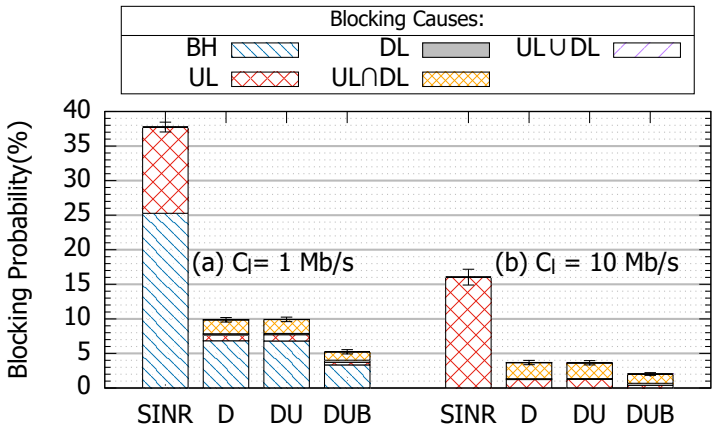


Figure 4.6: Blocking causes in best SINR and DUB for (a) C_l = 1 Mb/s, (b) C_l = 10 Mb/s.

specifically the UL, that creates a bottleneck. One could argue that this could be a consequence of the assumption we made in the input data that the UL traffic is symmetric to the DL traffic. However, this is reasonable in many applications, such as video calls [29]. Moreover, even with asymmetric traffic scenarios, not shown here, observations are the same. Indeed, interference being more aggressive on the UL than on the DL, for the same throughput, more resources are needed on the UL than on the DL, causing the UL to saturate faster. On the other hand, looking at the blocking causes in the DUB policy, we notice that the bottleneck effect is mitigated, and the UL is not a major problem anymore. This shows that DUB, by trying to minimize the remaining UL resources at each association decision, allows a better utilization of available resources.

We decrease the capacity of the backhaul links, such that C_l = 1 Mb/s ; $\forall\ l \in \mathcal{L}$ (Fig. 4.6, case (a)). The network is now backhaul-limited, since, practically, a link can only carry one flow at a time. Results show that the high blocking probability in the best SINR policy is mainly due to the backhaul. By reducing the link capacities, the number of feasible paths between sdNodeBs is also reduced, saturating the backhaul and creating a supplementary bottleneck. Nevertheless, the DUB policy overcomes these limitations by taking into account the remaining backhaul resources. Indeed, the flow blocking probability is reduced by a factor of 9.

These results show that, by ignoring the backhaul, current association policies are prone to underperform. This is problematic in self-deployable scenarios, where the backhaul is not only limited, but also dynamic, since the sdNodeBs are

mobile. Being able to estimate and predict the state of the backhaul is essential in this case, allowing significant gains in terms of flow blocking probability.

4.2 Mobility Management in CRAN Architectures

The increase and diversification of user demand puts an increased pressure on the RAN, which must cope with an ever-larger number of users presenting heterogeneous traffic demands. Mobile operators are addressing this issue by developing more flexible and efficient RAN mechanisms, as well as through a densification of the access network. However, adding more BSs comes at the price of increased interference and important financial costs: antennas, also known as remote radio heads (RRHs), need to be installed on a high point and connected to a baseband processing unit (BBU) nearby. A growing problem is the fact that the high computational load of the BBU requires the installation of a cooling system, increasing the energy consumption and further limiting the locations where a BS can be installed.

The self-deployable approach discussed in the previous section is one solution to the problem, but not yet ready for massive adoption. A more reasonable idea at the present time is the use of a centralized RAN [45], where only RRHs are installed on site, and connected through an optical fiber link with their BBUs, which are instead gathered in a data center. This allows a larger choice for the installation sites and simplifies the deployment.

With the parallel development of virtualization techniques, the Cloud RAN (CRAN) architecture emerged [86]. In CRAN, radio access functions such as signal processing, resource allocation or mobility management become software functions, running on any of the available computational resources [17]. The one-to-one mapping between RRH and BBU is therefore no longer needed. Multiple BBUs are implemented in a BBU pool and one BBU can handle several RRHs at a time. Moreover, since the load generated by an RRH varies with time, a dynamic mapping between RRHs and BBUs can further improve the usage of the computational resources [50].

Several major aspects of the CRAN architecture have been addressed in the last few years: the possible cost and energy savings [9], the limits imposed by the fronthaul connecting the antenna site and the data center [72], or the interference appearing between RRHs [79]. However, mobility management in this new architecture has been disregarded, despite its impact on user experience and the fact that CRAN highly modifies the current network mechanisms. Indeed,

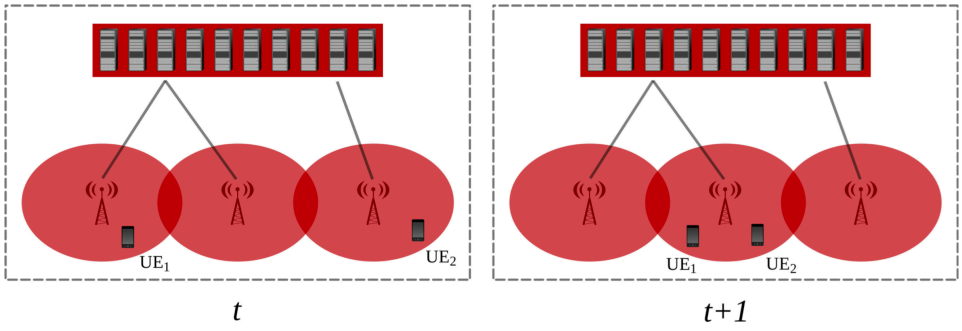


Figure 4.7: Illustration of an MHO occurring between successive snapshots t and $t+1$. The mobile user UE_1 does not encounter an MHO, while UE_2 does.

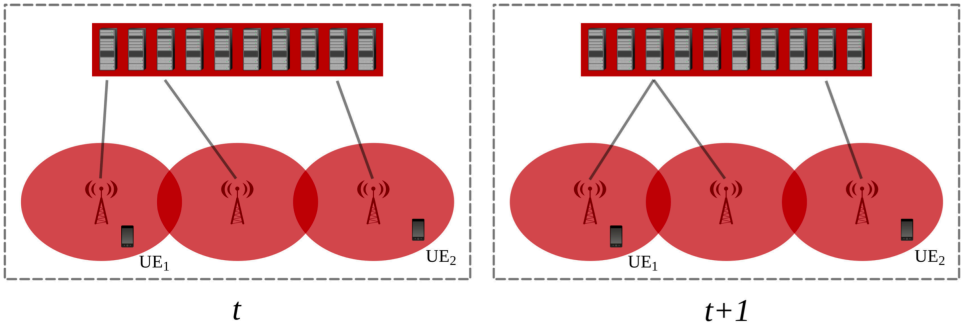


Figure 4.8: Illustration of an RHO at successive snapshots t and $t+1$. Static user UE_1 encounters an RHO, while static user UE_2 does not.

in CRAN, UEs are associated on the RAN with a BBU, which can serve several RRHs. This means that an UE movement between two different RRHs does no longer result in a mobility handover (MHO), if both RRHs are handled by the same BBU, as illustrated in Fig. 4.7. While this indicates a reduction in the number of handovers, seemingly improving the user experience, we notice a new type of handover, specific to the CRAN environment and related to the dynamic mapping between BBU and RRH: when an RRH changes its BBU, all the users covered by that RRH have to change their BBU association as well, as shown in Fig. 4.8. This reconfiguration handover (RHO) was previously undocumented in the literature and the question in this case is whether the overall number of handovers in the system decreases or not.

The number of handovers in a mobile network is directly related to the user experience. During the handover procedure, ongoing UE packet flows can be either re-routed to the new BBU or lost. In both cases, the communication is disrupted, impacting the quality of experience (QoE) perceived by the UE. In fact, handovers have been observed to lead to a 10% increase in video session

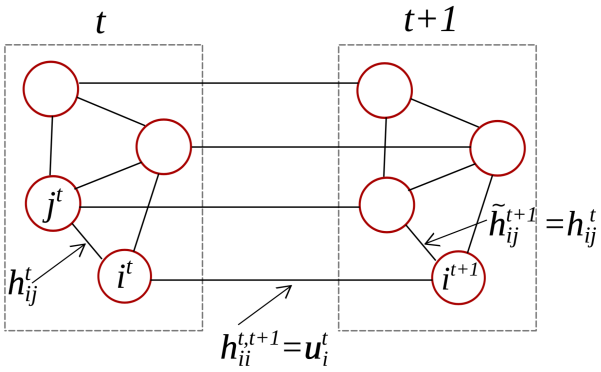


Figure 4.9: Time-varying graph representation of a CRAN architecture.

abandonment rates [66]. Web QoE has also been shown to be affected by handovers, with most web sessions abandonned in presence of handovers [7]. IP-level measurements confirm these results as well, showing disconnections that can reach tens of seconds [37]. Accordingly, handovers play a critical role on customers satisfaction and need to be properly handled.

4.2.1 Time-Varying Graph Representation

To study this handover problem, we consider the set of snapshots $\mathbb{T}$ to represent the access network traffic, with each snapshot t providing information about UEs demand over a time interval ΔT_s. ΔT_s represents the smallest scheduling time interval for a particular cellular network standard, i.e. 1 ms in LTE systems. Each couple of consecutive snapshots that we consider in $\mathbb{T}$ are separated by a time interval ΔT_r, with $\Delta T_s \leq \Delta T_r$. ΔT_r refers to the smallest time interval for CRAN reconfiguration, i.e. the smallest time interval possible for a BBU-RRH association.

We use i^t to denote an RRH i in the network during snapshot t, and $\mathcal{R}^t$ to represent all RRHs over t. We refer to the complete set of RRHs over $\mathbb{T}$ as $\mathcal{R}$. We then construct a time-varying graph structure $G(\mathcal{R}, \mathcal{E})$, over the set $\mathcal{R}$, as illustrated in Fig. 4.9.

The set $\mathcal{E} = \{\mathcal{E}_m \bigcup \mathcal{E}_r\}$ is the set of edges, representing the set of candidates MHO $\mathcal{E}_m$, as well as the set of candidates RHO $\mathcal{E}_r$. The set $\mathcal{E}_m$ is formed by subsets $\mathcal{E}_m^t$, such that an edge $e_{ij}^t \in \mathcal{E}_m^t$ exists between i^t and j^t if the cells corresponding to RRHs i and j in the Voronoi diagram of the access network share a common border. Each edge e_{ij}^t is assigned a weight h_{ij}^t referring to the

total number of active users moving between RRHs i and j over the time interval ΔT_r, spanning between $t-1$ and t.

The set $\mathcal{E}_r$ is also formed by subsets $\mathcal{E}_r^{t,t+1}$, such that an edge $e_{ii}^{t,t+1} \in \mathcal{E}_r^{t,t+1}$ links node i^t to node i^{t+1}. We assign a weight $h_{ii}^{t,t+1} = u_i^t$ to edge $e_{ii}^{t,t+1}$. u_i^t refers to the total number of users connected to RRH i^t and thus represents the total number of potential RHO, which would be triggered in case RRH i^t and RRH i^{t+1} would be associated to different BBUs. In the following, we consider that the u_i^t users associated with RRH i^t require d_i^t resources during the time interval ΔT_s covered by snapshot t. At each snapshot, each RRH is associated to a BBU k, which we consider to have a fixed capacity of resources f_k.

Our goal is to find an association between RRHs and BBUs that minimizes the number of handovers in the system. Using the time-varying graph representation, this is equivalent to grouping the graph nodes (i.e. the RRHs) into a set of clusters (i.e. BBUs), such as to minimize the weight of the edges (i.e. the number of handovers) connecting different clusters. We also need to take into account that each RRH has a resource demand, related to the number of UEs it serves, and that the BBUs have a limited resource capacity. Therefore, we formally define our problem as follows:

$$\min\left(\sum_{i^t\in\mathbb{C}_k, j^t\in\mathbb{C}_l} h_{ij}^t + \sum_{i^t\in\mathbb{C}_k, j^t\in\mathbb{C}_l} h_{ii}^{t,t+1}\right)$$
$$s.t. \sum_{i^t\in\mathbb{C}_k} d_i^t \leq f_k \quad \forall k, \forall t \tag{4.1}$$

If we do not consider the supplementary capacity constraint in Eq.(4.1), our problem is a classical graph clustering, or community detection problem [32]. A number of techniques have been proposed for the community detection problem [47], which is known to be NP-hard [12]. One of the most popular heuristics is the Louvain method [11], known for its scalability and good performance. The Louvain method optimizes a metric known as *modularity*, with values ranging between -1 and 1. Practically, the modularity of a graph partitioning compares the cohesion inside partitions to the case of a random distribution of edges over partitions. A high value of modularity indicates a high cohesion of links inside each partition, with respect to the links among them.

However, the Louvain method only accounts for the links inside each partition, and could theoretically results in very large partitions, not realistic in our case since a BBU can only support a limited number of RRHs simultaneously. We

therefore adapt the Louvain method to include the capacity constraint and, based on this, propose two solutions to our handover problem. The first is an oracle-based solution, where we consider we know in advance the traffic demand and the user mobility for the next 24h. In this case, we find the RRH-BBU mapping that gives the best clustering in the time-varying graph described above, i.e. it minimizes the number of handovers in the network. The second solution is an online one, without a priori knowledge of the network state. In this case, we need to predict two important KPIs: the number of UE movements $\widetilde{h}_{ij}^{t+1}$ taking place between RRHs i and j over the coming ΔT_r period and the demand of users $\widetilde{d}_i^{t+1}$ over RRH i^{t+1}. As in the case of Sec. 4.1, we do not put a lot of energy in investigating prediction methods; we simply make the assumption here that $\widetilde{h}_{ij}^{t+1} = h_{ij}^t$ and $\widetilde{d}_i^{t+1} = d_i^t$. Practically, we thus assume that the mobility of users and their traffic demand across the network does not encounter important variations over ΔT_r. While this prediction might not be very accurate, depending on the value of ΔT_r, it already provides very good results, as shown below.

4.2.2 RRH-BBU Mapping Evaluation

We evaluate the oracle and online strategies in two scenarios, using the mobile traffic datasets provided in the context of the D4D challenges [10, 22], and covering the urban areas of Abidjan, in Ivory Coast, and Dakar, in Senegal. However, the two datasets do not provide any information regarding user mobility. Therefore, we derive this by assuming that x% of the total communications experience a handover. We then randomly pick the UEs suffering a handover and choose the destination cell according to a uniform distribution over the set of neighboring cells. To test the impact of user mobility on our RRH-BBU mapping solutions, we test two use-cases, with the assumptions of 5% and 50% of total calls encountering a handover, representative of two extreme cases: low and high mobility scenarios.

We compare our solutions to two BBU-RRH mapping strategies. The first one aims at optimizing the mapping between BBUs and RRHs from a frequency utilization perspective. More precisely, its objective is to efficiently utilize frequency resources, in order to minimize the interference in the system. Instead, the second strategy aims at optimizing BBU-RRH mapping from a capacity utilization perspective. In particular, it aims at minimizing the number of BBUs that are being used in the network. Both these strategies operate over separate individual traffic snapshots. We briefly describe them in the following.

Mobility \ Strategy	Oracle	Frequency	Capacity	Online
Low	0.77	4.79	4.4	0.77
High	0.76	1.35	0.95	0.71

Table 4.1: Ratio of handovers R for different RRH-BBU mapping strategies in Abidjan.

Mobility \ Strategy	Oracle	Frequency	Capacity	Online
Low	0.85	3.06	2.74	0.78
High	0.79	1.19	0.85	0.79

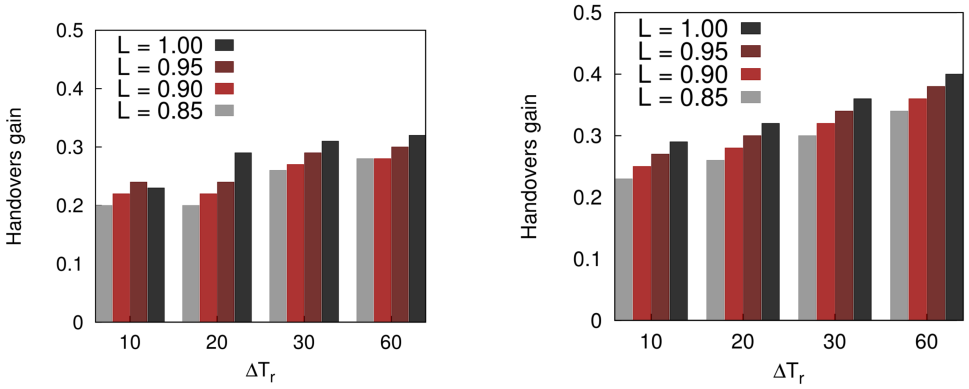
Table 4.2: Ratio of handovers R for different RRH-BBU mapping strategies in Dakar.

Frequency-oriented strategy. We employ the frequency-oriented BBU-RRH mapping algorithm proposed by Wang *et al.* [79]. The algorithm determines the mapping based on a dynamic frequency reuse scheme. It applies a graph coloring method over separate snapshot graphs. In each graph, nodes represent RRHs and edges link neighboring RRHs. As such, the method does not allow to use the same color, i.e. the same frequency, for adjacent nodes.

Capacity-oriented strategy. This algorithm determines the mapping between BBUs and RRHs with the objective of minimizing the number of BBUs. It employs a greedy approach over a traffic snapshot. The algorithm first selects the RRH with the highest user demand and places it on a BBU. It then goes through the list of its neighbors, by decreasing order of demand and places them over the same BBU, as long as its capacity is not violated. The algorithm then covers all nodes in the network by considering multi-hop neighborhoods, one after the other. Once a BBU reaches its capacity limit, another BBU is instantiated.

We compute the ratio $R = \frac{HO_{CRAN}}{HO_{RAN}}$, representing the ratio between the number of handovers in CRAN obtained for each strategy HO_{CRAN} and the number of handovers existing in a traditional RAN HO_{RAN}. The corresponding results are presented in Tab. 4.1 and Tab. 4.2, for each dataset, in the two mobility scenarios.

Several observations can be made based on these results. First of all, both frequency- and capacity-oriented strategies lead to ratios higher than 1, especially in low mobility scenarios. This means that these RRH-BBU mapping strategies, while optimizing other metrics, would increase (sometimes multiplying by 4) the number of handovers experienced by users.



(a) Low mobility.

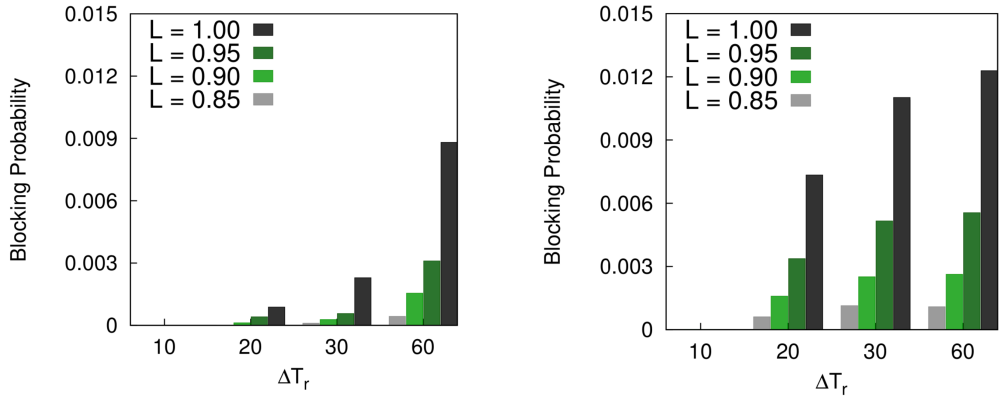
(b) High mobility.

Figure 4.10: Handovers gain for different ΔT_r values and capacity thresholds L, when using the online strategy.

Second, both the oracle and online approaches that we propose reduce the number of handovers with respect to a classical architecture, by around 20%, despite having to consider a second type of handover, namely RHO. Delving deeper in the results (not shown here), we observe that this is achieved by using slightly more BBUs: 10 supplementary BBUs are required at peak hour by our solutions when compared with the capacity-oriented strategy, while reducing the handovers between 20% and 400%.

Finally, we note that the online method can lead, in some cases, to a lower number of handovers with respect to the oracle strategy. This surprising behavior can be explained by the fact that the online strategy makes, as explained, a prediction of future traffic in snapshot $t + 1$. As a result, when the UE demand is underestimated, the BBU capacity can be inferior to the number of required resources, producing a call drop. These call drops are not allowed in the oracle strategy and the apparent gain of the online strategy is the consequence of this extra degree of liberty. To further investigate this phenomenon, we present supplementary results related to the Abidjan dataset in Fig. 4.10 and Fig. 4.11 (the results for the Dakar dataset are very similar).

A parameter with an important impact on the quality of the estimation is the CRAN reconfiguration time ΔT_r: as this value increases, the time window that needs to be estimated increases as well, and the quality of the estimation decreases. This leads to blocking more calls as a result of higher errors in future traffic estimation. A second parameter with significant consequences on the online algorithm is the access control capacity threshold L. Practically, for BBU



(a) Low mobility.

(b) High mobility.

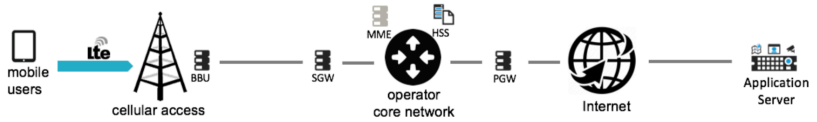
Figure 4.11: Blocking probability for different values of ΔT_r and L, when using the online strategy.

k with capacity f_k, the access control policy allows an RRH to map to k only if the total demand of the BBU does not exceed $L \cdot f_k$.

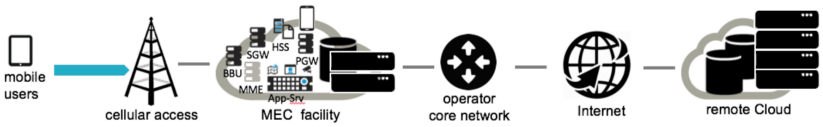
In Fig. 4.10, we show the handover gain, defined as $1 - R$, for the online algorithm with different values for parameters ΔT_r and L. We plot the obtained gain for ΔT_r equal to 10, 20, 30 and 60 minutes, and different capacity thresholds L of 0.85, 0.9, 0.95, and 1, in the low and high mobility scenarios. The figure shows that, as higher values of ΔT_r and L are considered, we generally get higher gains in terms of handovers. In fact, higher values of ΔT_r translate into less reconfigurations in the system, meaning less RHOs and more MHOs.

However, as shown by Fig. 4.11, this comes with the price of an increasing call blocking probability because of exceeding the capacity limit of a BBU. We can notice that, in the extreme case of $\Delta T_r = 60$ minutes and $L = 1$, the percentage of blocked calls is as high as 1.5%, which is not acceptable for mobile operators. Therefore, the price to pay for saving 40% of handovers might be too important. However, even a small tolerance for call blocking probability, in the order of 10^{-4}, can result in handover gains of more than 30%, for example when $\Delta T_r = 60$ minutes and $L = 0.85$.

In summary, these results indicate that using the online strategy, with adequate parameters, leads to important savings in terms of handovers in the network, offers a more efficient management of BBUs, and grants a lower system reconfiguration frequency. Of course, the results could certainly be improved by using a better prediction algorithm, but even our naive approach shows well enough the gains obtained by an anticipatory networking solution.



(a) Legacy 4G mobile network.



(b) Fully cloudified mobile network.

Figure 4.12: Mobile access network evolution with edge computing.

4.3 Mobile Edge Computing Orchestration

The RAN is not the only part of a mobile network that can benefit from advances in virtualization. Indeed, the type of functions that can be virtualized ranges from traffic load balancers and multimedia (de)coders to mobile core functions such as those of the LTE EPC [42]. Application servers can also be run in such facilities, so that the end-to-end user experience benefits from low access latency [3].

An illustration of this evolution is given in Fig. 4.12. Fig. 4.12a depicts a legacy 4G environment, where the user accesses remote applications via cellular access, in such a way that its wireless signals are processed at BBU nodes integrated to cellular BSs, its traffic is routed through the EPC (composed of four main functions: SGW, PGW, MME, HSS), before reaching the Internet border on the way to the application server. Fig. 4.12b shows instead a fully cloudified environment, where radio-network elements such as the BBU, EPC functions, mobile phone remotely executable applications, as well as application servers (possibly synchronized with a remote cloud) are all virtualized in potentially the same place, called Mobile Edge Computing (MEC) facility. Such a scenario is an extreme one, coping with the virtualization of a variate set of hardware, but that could correspond to the reality in the coming decade.

Among the virtualizable nodes at MEC facilities, we can distinguish nodes that are strictly serving a subset of the BSs of an operator (e.g., vBBU and vEPC nodes), and nodes that serve single or multiple users (e.g., virtualized mobile device environment for computation offloading, virtualized application servers), possibly behind different BSs. An important amount of traffic can therefore be aggregated at MEC facilities, depending on the type of virtualized functions

that are run at these edge delivery points. The management of virtualized nodes running at MEC facilities encompasses service and network management operations mainly related to: *i)* BS-to-MEC facility association, and *ii)* user-to-virtual machine (VM) association (a VM being in turn associated to a MEC facility).

At the time being, the telco industry is more focused on the virtualization of the nodes that serve a subset of cellular antennas (e.g., vBBU, vEPC), instead of working directly at the user-VM granularity, mainly because of scalability concerns. Therefore, one shall consider BS-to-MEC facility switching decisions as critical ones. In the following, we propose a MEC orchestration framework that primarily optimizes BS-to-MEC facility association over time, based on a spatiotemporal grouping of the BSs, while integrating VM workload adaptations across MEC facilities.

BS-to-MEC facility switching operations can not be reasonably expected to run continuously, as this would incur in traffic loss and overhead due to traffic handover, but to occur only at certain points in time (*e.g.*, once every thirty minutes). Hence, introducing an implicit time discretization of the orchestration system appears appropriate. For this, we use the temporal profiling approach presented in Sec. 3.1.

4.3.1 Network Orchestration Model

We elaborate our reference MEC network orchestration model along the following generic lines. BSs have associated mobile traffic demand, that changes over time. Each MEC facility has a certain capacity, limiting the overall amount of demand it can serve simultaneously. BSs must be associated to MEC facilities. A user connecting to the network (through a BS) pays an assignment cost to reach the corresponding MEC facility. Since MEC facilities have a finite capacity, BSs are not necessarily associated to the MEC facility of minimum latency. Furthermore, since demand changes over time, an assignment pattern would hardly remain an efficient one over the whole planning horizon. We therefore leave the option of changing assignments (and BS associations) over time, taking into account that each change implies a switching cost for the network, for example in terms of signaling to move session data of active users. An optimization problem therefore arises, that is to associate BSs to MEC facilities over time, respecting capacity constraints and minimizing a combination of users (assignment) and network (switching) costs.

This optimization problem is solved by an orchestrator, using the approach described in [16]. In particular, the orchestrator builds dynamic assignment plans detailing, for each time slot, the set of BSs to be associated to each MEC facility and, as a by-product, the set of switching operations to be performed between subsequent time slots. We consider a periodic single-association operational policy, that is, in each time slot each BS is associated to exactly one MEC facility, and the last time slot is assumed to be followed by the first one.

The problem details are the following.

Input. We assume to be given the set of BSs, the set of MEC facilities and a discretization of the time horizon in a set $\mathbb{T}$ of time slots. We also assume to be given *i)* for each BS, the mobile traffic demand that has to be accommodated in each time slot, *ii)* the capacity of each MEC facility, *iii)* the physical distance between each BS and each MEC facility and the network distance between each pair of MEC facilities (that is, a measure directly proportional to the network latency, including packet processing latency at intermediate nodes, and physical distance).

Output. We expect, as output of the optimization problem, an assignment plan: for each BS and each time slot, an indication of the MEC facility where traffic needs to be routed. As a side result, we expect a switching plan, that is a boolean value for each BS and each pair of MEC facilities for each time slot, indicating whether that BS switches at that time between a particular pair of MEC facilities, or not.

Requirements. The assignment plan satisfies the following conditions: *i)* the overall demand assigned to each MEC facility at each time slot must not exceed its capacity, *ii)* each BS is connected to exactly one MEC facility at each time slot, *iii)* assignment and switching plans must be coherent.

Objective. The plans must target the minimization of network- (switching) and user-related (assignment) costs. The former is generated by the change of BS-MEC facility associations in consecutive time slots, which produces some overhead due to the necessity of migrating VMs, but also energy and bandwidth costs for the network. The latter is instead the latency experienced by a user connecting to the network with the current BS-MEC facility association. As we will see, these two terms have an opposite behavior, meaning that a trade-off between these two costs needs to be found.

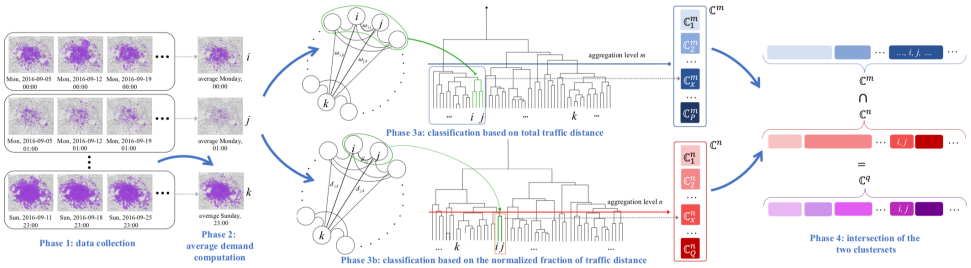


Figure 4.13: Workflow for the classification of network usage profiles. The final intersection cluster set is used for the training phase.

4.3.2 Data Analytics

In order to decide when should the orchestrator switch between association configurations, two approaches can be taken. The simplest option is to have a periodic decision, based on aggregated demand observed at each BS during a recent training period. Another option is to identify suitable discrete-time profiles of the traffic demand using some kind of mobile data analytics, so as to group together time slots that feature very similar distributions of the mobile traffic demand across the BSs.

The analytics required by this second approach are very close to the temporal classifier of mobile network traffic discussed in Sec. 3.1. However, we want to define our profiles based on both the overall traffic volume (since the MEC facilities have a fixed capacity) and traffic distribution (since we want to produce a spatial mapping between BS and MEC facilities). The two metrics we used in Sec. 3.1, $\mathcal{V}$ and $\mathcal{D}$, account for one or another of these constraints. Therefore, we slightly adapt our profiling work to this scenario and combine $\mathcal{V}$ and $\mathcal{D}$. This produces one single cluster set (unlike in Sec. 3.1, where we analyzed the two profiles individually), representing the intersection of the clusters produced by the two metrics. The resulting cluster set is very similar to the examples shown in Fig. 3.4, with the difference that we obtain more profiles. A sample cluster set obtained in this work is depicted in Fig. 4.14.

For evaluation, we use a dataset collected in the core network of Orange in the context of the ANR ABCD project, during three months in 2016. It describes the traffic generated by several millions of mobile subscribers in the French metropolitan areas of Lyon and Paris, for a specific mobile service, *i.e.*, Facebook, with a granularity of 10 minutes. The rationale for the choice of Facebook is that it represents a prominent mobile service, generating around

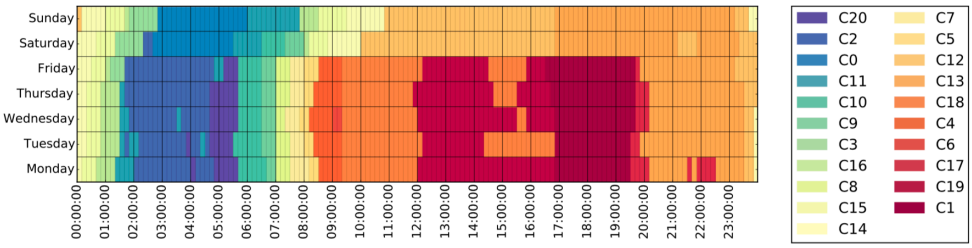


Figure 4.14: Sample cluster set of 10-minute time instants for a mobile service. The plot outlines the existence of 20 temporal classes of spatial distributions of the demand, during each daytime (abscissa) of different weekdays (ordinate).

20% of the compound downlink and uplink demand in the network. It is also an example of typical service that could benefit from the improved quality of service granted by a MEC infrastructure. The Lyon dataset contains demands from 332 BSs, while the Paris dataset has 1907 BSs. We set three cardinalities of facilities for the dataset of Lyon (10, 20 and 30, resp.) and two cardinalities of facilities for the dataset of Paris (20 and 50, resp.). The location of the facilities was generated by a k-medoid algorithm, using the coordinates of the BS locations as input data. For the training set, we use the first 4 weeks of the dataset to build the temporal network profiles.

4.3.3 MEC Orchestration Evaluation

We compare seven different solutions below. As benchmark, we considered a baseline approach without any temporal aggregation, therefore with a single time-period, leading to a single association for every BS to a MEC facility over the week and no switching among MEC facilities during the week. This static scheme is denoted as 'S' in the figures.

We also evaluate six different time-period aggregations:

- a static aggregation over four, two and one hour ('4H', '2H' and '1H' in the remainder). The resulting training sets consist of 42, 84 and 168 time-periods, respectively. Using shorter time periods for this strategy revealed to be too complex to solve.

- an aggregation over the temporal profiles produced by our framework, with clustersets of one hour ('1HC'), 30-minute ('30MC') and 10-minute ('10MC'). The resulting training sets differs for Lyon and Paris dataset: for

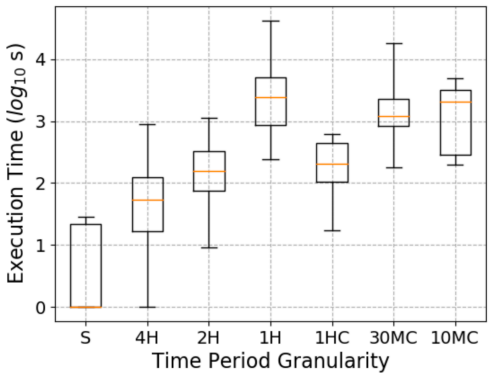


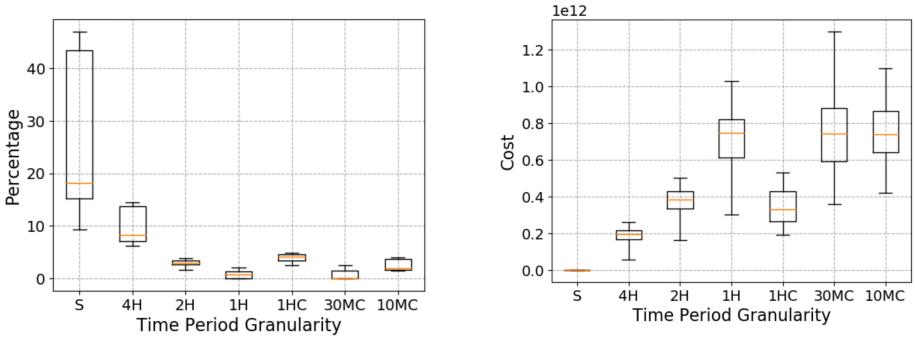
Figure 4.15: Box-plots of execution time for the training phase (please notice the log axis).

Lyon dataset, the number of profiles is 99 for '1HC', 171 for '30MC' and 260 for '10MC'; for Paris dataset, we obtain 60 for '1HC', 160 for '30MC' and 141 for '10MC'.

In the following, we represent results as boxplots, where the box bars indicate the minimum, 1st quartile, median, 3rd quartile and maximum of the plotted metric. We first show, in Fig. 4.15, the execution times of the training sets, in logarithmic scale. We can notice that the consecutive 1H case has the highest execution time (up to 10 hours), followed by the 30MC case. Practically, what we notice here is that the approaches with a similar number of profiles, e.g., '2H' and '1HC' (resp. '1H' and '30MC'), require similar training time.

Next, we use the remaining 8 weeks of data to test the assignments generated by the orchestration algorithm in [16] on a weekly basis. That is, for each approach, we have eight tests, with different demands, as given by the Orange dataset. We present two types of costs in Fig. 4.16: *i)* the assignment cost considering the distance between a BS and MEC facility and the demand of the BS, and *ii)* the switching cost considering the distance between MEC facilities in consecutive periods and the demand of the BS. We show the two costs separately, because have different units and they are difficult to combine in a single total cost. Indeed, the assignment cost is given by the latency perceived by the users when connecting to the network. The switching cost is more complex, including a latency cost on the user side, but also an energy and bandwidth cost for the network.

In the case of the assignment phase, in Fig. 4.16a, we show normalized results, as a percentage of extra-cost with respect to the best case among those tested. The results show that the static approach 'S' always leads to the highest



(a) Assignment cost. (b) Switching cost.

Figure 4.16: MEC assignment and switching costs gaps.

cost, on median 20% higher than the rest. This proves the interest of switching assignments, since all the other solutions show much similar costs, all similar, except '4H' which has slightly worse results (on median 8% higher cost).

For the switching part, we can not use a similar normalization, since the static approach 'S' does not require any switching. We therefore show absolute values, as obtained when solving the switching optimization problem in [16]. We note that this cost does not have any physical meaning, and it is just an artifact of our model. In this case, '4H' always leads to the lowest cost, i.e., the least number of MEC facility switching, followed by '1HC' and '2H'.

Considering the differences in the two costs, proposing a total cost representative of the trade-off between assignment and switching seems difficult. We could probably argue that, considering the training time and the two costs, the '1HC' solution is the best choice for the MEC orchestration problem, as it gives a relatively low assignment cost, while producing at the same time one of the lowest switching costs. However, it is fair to say that this study did not really show a major advantage for the temporal profiling solution we proposed. But the results still clearly demonstrate the importance of an anticipatory approach: the six switching-based solutions use temporal classification and prediction (whether over regular intervals or over temporal profiles) and reduce the assignment cost (directly related to the access latency) by 10% to 50% with respect to a static approach.

*****

While I appreciated working on data collection and data analytics, the best moments during the last seven years have always been those when I was working on some networking questions. Describing a previously undocumented type

of handover or designing from scratch a self-deployable network architecture represent experiences that I am particularly fond of.

One evolution I noticed in my research on networking subjects comes from my increased interest in cellular networks, or in operated networks in general. This is a direct consequence of two factors. First of all, I was coming after a PhD in vehicular ad-hoc networks, where cars would just meet and exchange information. Following this field was, and still remains, a frustrating experience for me: while technical solutions for vehicular communications exist, they are yet to be largely implemented and deployed, mainly for political reasons[2]. Second, I gradually reached the conclusion that running a network is a (very) difficult job. Deploying objects around and letting them spontaneously communicate with one another is a beautiful idea, but a technical nightmare.

However, this does not mean that my research object has become the cellular networks operated by a handful of operators. My vision is one where network operation is democratized, where a cellular network is based on flexible, lightweight mechanisms, such that running it becomes a simple task, approachable even by non-experts.

*****

This chapter contains results published in the following articles:

1. Jad Oueis, Catherine Rosenberg, Razvan Stanica, and Fabrice Valois, **Network-aware User Association in Public Safety Oriented Mobile Networks**, ACM 1st International Workshop on ICT Tools for Emergency Networks and Disaster Relief (I-TENDER 2017), Incheon, December 2017.

2. Diala Naboulsi, Assia Mermouri, Razvan Stanica, Hervé Rivano, and Marco Fiore, **On User Mobility in Dynamic Cloud Radio Access Networks**, IEEE 37th Annual International Conference on Computer Communications (INFOCOM 2018), Honolulu, April 2018.

3. Alberto Ceselli, Marco Fiore, Angelo Furno, Marco Premoli, Stefano Secci, and Razvan Stanica, **Prescriptive Analytics for MEC Orchestration**, IFIP 17th International Conference on Networking (NETWORKING 2018), Zurich, May 2018.

[2]To put it simply, there is no organization interested in lobbying for these solutions to become mandatory in new (and old) cars.

Chapter 5

The Last Stage

Also known as Perspectives

Roads go ever ever on,
Over rock and under tree,
By caves where never sun has shone,
By streams that never find the sea.

The Hobbit, or, There and Back Again
J.R.R. Tolkien (1937)

I tried to summarize in these pages my work during the last seven years in the field of wireless mobile network. From an administrative point of view, this document is meant to demonstrate to the reader that I am a researcher worthy of teaching others. From a personal point of view, these chapters tell my story, of my growing up, learning from others and calming down (but not completely) throughout the years. From a scientific point of view, this manuscript presents my vision of the networking field, pushing for a more data-driven approach in our field.

Networking is still a young field, with a history of just 50 years. With this in mind, we need to look around and take inspiration from topics with an older history. It seems quite obvious to us today that roads are built differently depending on the type of vehicle they are meant for. No one argues against the idea of building pipelines differently depending on what they need to transport. Depending on the type of energy they transport, power lines are not constituted in the same way. But computer networks are still agnostic to the data they transport, designed to transport video and email and small data, all mixed together.

My interest in anticipatory networking solutions comes directly from this diversification of usages and it relies on the idea that a network, in order to function at its best, needs to adapt to the context of the user and to the content that it needs to transport to/from the user. This can only be done by observing continuously the state of the network, by measuring well-selected KPIs, and by analyzing this collected data in order to understand and predict the behavior of the network.

I will finish this story by discussing some perspectives, general for the networking field, and personal, regarding my future research.

5.1 General Perspectives

The goal of this section is not to discuss the evolution of 6G or that of the next generation of WiFi, but to address three important factors that I believe will shape wireless networks in the following years. The first factor I would like to mention is softwarization, and its little brother, virtualization. SDN is no longer a research-only concept, it is an approach well established in wired networks. Since wireless networks need specific equipment, whether cellular base stations, WiFi access points, or other antenna-based hardware, it escaped the first SDN wave. However, this started changing in the last few years, with open software projects, such as OpenAirInterface [58] or srsLTE [39], implementing the entire cellular network stack, which can now be run on a normal computer. With network software easily executable on any platform, functions can be easily added and removed, or they can be adapted to different types of data.

A second factor I see emerging stems from the first: network automation. One could argue that network automation has been announced as the next big step for the last two decades, with a questionable success. But the only automation we can achieve today is through some scripts that exploit weird regular expressions in order to generate some access lists or to set up a new equipment. When network functions will be able to run anywhere, their execution will need to be automated. And, as much as some of us might dislike it, this will be the perfect playground for machine learning: technically-complicated tasks, that only a few humans understand, but where training data is abundant.

Finally, the third factor I would like to discuss is a societal one, and regards privacy and ethics. For now, the discussions around these subjects mostly focus on application-level issues: social networks influencing us, targeted advertisement, biased decision-making algorithms and so on. These problems remain largely

neglected by a majority of the population, but a vocal minority might be able to educate the masses and create a political agenda around these subjects. However, if the society shifts towards privacy-preserving applications and ethical algorithms, one needs to expect networking to follow. This would not only require major evolutions in networking architectures, but also a societal implication in a field where this is not common.

5.2 Short Term Perspectives

Since I structured my contributions in this document around three major topics, I feel it is suitable for a conclusion to discuss some perspectives with respect to these topics. As these are short term perspectives, these are practically ongoing works, with some preliminary results already available.

5.2.1 Data collection

Regarding data collection, the missing piece today is the compliance with privacy-protection laws. The challenge is not necessarily to collect data, but to publish and share this data in an anonymized way. Mobile traffic data proved useful in many research fields and granting access to these massive datasets could bring major scientific advancements. However, sharing such datasets is not possible without good anonymization solutions. In this sense, the definition of anonymized data needs to be clarified first. Then, from a scientific point of view, anonymization algorithms for behavioral and service consumption data need to be designed and their impact on the quality of the data has to be evaluated.

For us, users of such datasets, the most important question is whether anonymized data will remain useful for our applications. In this sense, I believe the foremost reason why pseudonymisation does not work is the high unicity that affects mobile phone trajectories. Unicity stems from the fact that mobile subscribers have very distinctive movement patterns, which make them univocally recognizable even in very large populations. Previous experiments showed that 50% of the mobile phone trajectories in a 25 million-strong dataset turned out to be unique when considering the three most frequent locations they contain [82]. Similarly, any mobile phone trajectory could be pinpointed with near certainty among 1.5 millions entries by just knowing five of its spatiotemporal points picked at random [21].

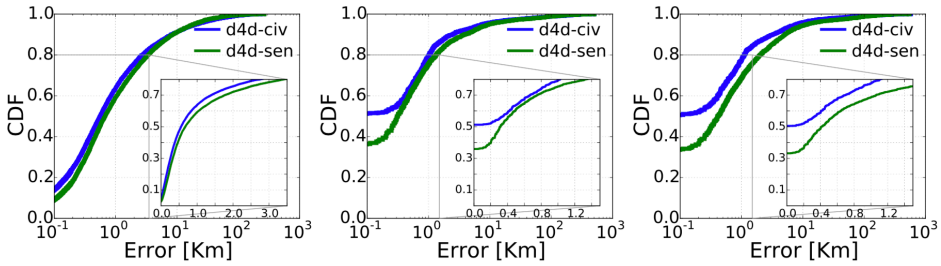


Figure 5.1: Estimation error of center of mass (left), home (center) and work (right) locations with 2-anonymized `d4d-civ` and `d4d-sen` datasets.

Mitigating unicity becomes then a very desirable facility that can entail more privacy-preserving datasets, and favor their publication and sharing. But publishing a highly altered private dataset would be of no use. We are therefore planning to assess the impact of different anonymization solutions on the quality of the produced data. We already have some preliminary results using a state of the art solution for mobile data generalization, namely GLOVE [40], which implements the k-anonymity privacy criterion in datasets of mobile phone trajectories[1].

To assess the utility of GLOVE-anonymized data, we used the two datasets released in the D4D challenges: Ivory Coast 2013 [10] and Senegal 2015 [22], denoted as `d4d-civ` and `d4d-sen`, respectively. We compare several classical mobility-related metrics, for the original datasets and their 2-anonymized versions obtained by GLOVE.

We first observe the center of mass, which denotes the pivotal location of a trajectory in space. The left plot in Fig. 5.1 shows the cumulative distribution function (CDF) of the error incurred when the center of mass is computed from 2-anonymized data with respect to the case where the same metric is derived from the original data. The error is below 0.5 km in around 50% of cases, and under 3 km in 80% of trajectories. If we consider that studies on human mobility [20] indicate that an accuracy below 7 km allows to model correctly the majority of people movements, almost all the centers of mass in the anonymized dataset are thus located correctly enough.

Home and work locations are important points of interest (PoI) in human mobility, classically derived as the most popular positions within a trajectory overnight and during working hours, respectively. The center and right plots

[1] In a k-anonymized dataset, at least k users present the same mobile phone trajectory. The unicity property is therefore removed.

in Fig. 5.1 portray the CDF of the error induced by 2-anonymized data when inferring such important locations. Depending on the dataset, 35% to 50% of the home and work locations are unaffected by the anonymization. At least 70% of these locations are placed within 1 km of their original position, and the 7 km threshold of [20] is met by over 90% of the trajectories.

These preliminary results are promising, as the properties of the original dataset do not seem significantly altered. However, different metrics and anonymization solutions still need to be tested. Important questions also remain regarding the preservation of outliers in anonymized data, which seems in total opposition with the idea of removing unicity.

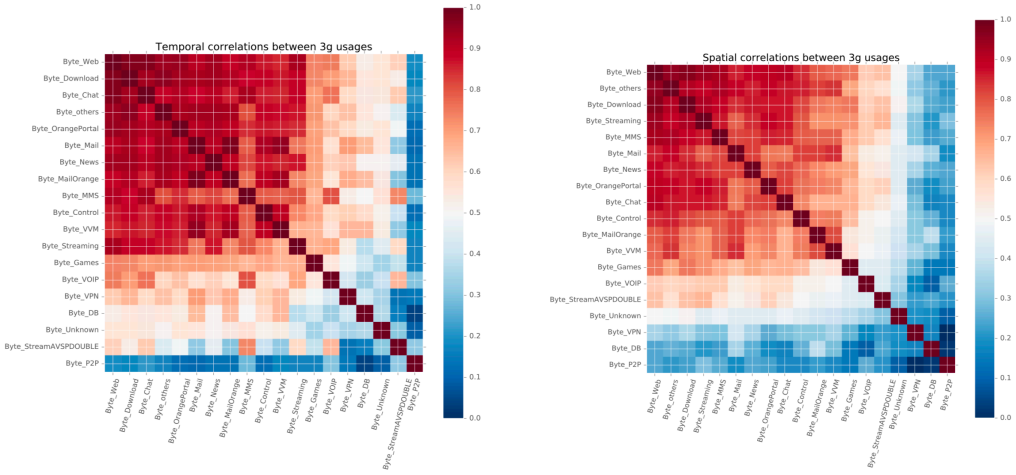
5.2.2 Data analytics

On the data analytics field, richer and richer datasets are becoming available. Most of the literature on mobile data analytics is based on CDR data which, despite the reduced temporal granularity, proved to be very useful in numerous fields. More recent datasets, collected from RNC and GGSN/SGW probes, provide not only call and text information, but also detailed downlink and uplink traffic data. With mobile applications heavily exchanging data in the background, the temporal granularity of these datasets is also more important.

Having access to per-service mobile traffic data can help refine the spatial and temporal profiling techniques detailed in Chapter 3. Fig. 5.2 shows preliminary results on the correlation between different classes of services, collected on the Orange 3G network in Lyon. While Fig. 5.2a focuses on the temporal correlation between services, Fig. 5.2b depicts their spatial correlation.

We can notice that, while some services are highly correlated (e.g. web traffic and file download), some others present very different patters (e.g. games). While using this data can bring new insight, it also raises an important challenge, by increasing the dimensionality of the problem. The so called *curse of dimensionality* is a well-known issue in data mining and machine learning [77], with no standard solution, which needs to be addressed in order to use such fine grained mobile traffic data.

Another stringent issue related to data analytics is the lack, for now, of online solutions. Practically, no operator in the world has a data analytics platform integrated with their data collection probes. Many proofs of concept have been developed on offline data, but adapting them to online streams of data raises important scalability problems. Nevertheless, if anticipatory networking is to



(a) Temporal correlation. (b) Spatial correlation.

Figure 5.2: Spatial and temporal correlation between different mobile traffic services over the 3G network in Lyon.

become a reality, classification, prediction and anomaly detection must run online and in real-time.

5.2.3 Anticipatory networking

Considering the anticipatory networking solutions, the most intriguing idea for me is the one of self-deployable networks. The fact that this new architecture tackles the status-quo in cellular network deployment makes it particularly interesting. Of course, a series of challenges need to be addressed before self-deployable networks become a reality. I have already addressed the user association problem, but our solutions in this area could certainly be refined. Scheduling the radio resources in the DL and UL is another important problem in cellular networks, which needs to be revisited in self-deployable scenarios. Finally, the integration of mobile base stations, carried by aerial or terrestrial robots, completely modifies the mobility management problem.

We have already started investigating this last point in the case of BS carried by unmanned aerial vehicles (UAV). In contrast to cellular networks, where the BS deployment is carefully planed and conducted, BSs mounted on UAVs are mobile themselves. Coordination among flying BSs and movement control are essential in this case, to set up a core network, provide the needed coverage, and ensure sufficient and stable user data rates.

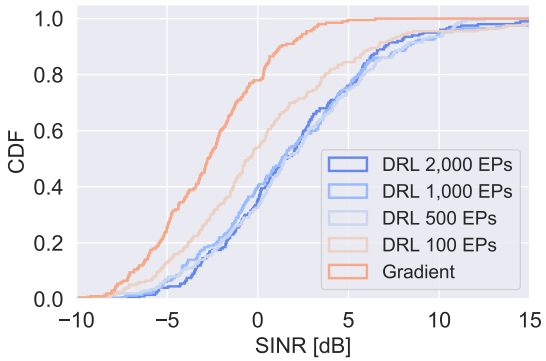


Figure 5.3: CDF of the SINR attained by all users with the proposed DRL over 10,000 testing steps, after 100, 500, 1,000, and 2,000 training episodes (EPs), and respectively with the benchmark gradient method.

In some preliminary results, we used a deep reinforcement learning (DRL) approach to tackle the challenges faced by UAV mobility control in this context. Practically, we devised a reward function that encourages the UAV mobility control agent to provide high quality signal coverage to users, and we leveraged an Asynchronous Advantage Actor-Critic (A3C) scheme to learn the optimal action policy via interaction with the wireless environment.

Simulation results demonstrate that our solution converges rapidly, and once trained, it makes accurate movement control decisions, outperforming a benchmark gradient-based scheme that has perfect knowledge of the stochastic channel, as shown in Fig. 5.3. More precisely, we obtain a 5dB median SINR improvement, while only requiring current location and association information.

However, these results only look at SINR, and more work is needed to understand the impact of this double mobility on the QoS perceived by users. Architectural challenges also appear, since the entire mobility management framework in cellular networks was designed with fixed BS in mind.

5.3 Long Term Perspectives

It is always difficult to talk about long term perspectives in research, even if they are not contractual. Writing this manuscript made this even more clear, as I would have definitely not imagined my research like this seven years ago. In fact, at the time I was planning to work on software-defined radio and multi-technologies networks. However, to conclude this document, I feel I need to play the game one more time, while adding a note in my agenda to come back and read this seven years from now.

5.3.1 Cellular network testbeds

The first subject I would like to develop is related to the experimentation of cellular networks. Classically, only mobile operators had access to cellular testbeds, where they could experiment with new algorithms and protocols. Academic research on the subject mainly focused on architectural issues, generally supported by analytic and simulation results. But studying the impact of different parameters at the radio resource and radio link control layers, or testing new mechanisms in realistic settings, remain tasks difficult to achieve, even for a large networking research team.

However, with RAN and CN software becoming available, I imagine a much more hands-on research on cellular-related topics. A democratization of cellular networks experimentation is already visible, with platforms such as OpenAirInterface [58] or srsLTE [39] gathering more and more users. These platforms will most likely become even more modular, allowing to easily activate and deactivate network functions.

In this context, my objective is to gather a fleet of terrestrial and aerial robots, each carrying RAN and CN functions, complemented by some fixed cellular infrastructure. This would be the perfect setting to test not only self-deployable solutions, but also problems related to network function placement, radio resource control, or quality of service.

One important theory I would like to test in this context stems from my belief that the separation of a cellular network in access and non-access strata is an outdated approach. Even in current deployments, RAN and CN equipment are often located in the same physical location[2], hinting at the possibility of a single stratum architecture. Designing, implementing and evaluating such a single stratum approach is a challenge I am looking forward to.

5.3.2 Network metrology

A second area I would like to gain expertise in is network metrology. My interest in this field is for now at the amateur stage, as I try to keep updated with the relevant literature and maintain the PrivaMov data collection application. But I plan to dedicate more time to the subject, and hopefully bring some contributions to the field. I can say that my attitude towards network metrology was, for a long time, presumptuous. My image was that of redirecting the output

[2]However, for all the operators I know, RAN and CN equipment are managed by different teams, leading to bureaucratic nightmares and numerous anecdotes.

of some system commands towards a log file. However, working on data collected by network probes, I became aware of the complexity of the task.

I can summarize two challenges I see in this field. First if all, a modern network equipment exposes hundreds of KPIs, covering all the layers, from the physical to the application one. Selecting and combining them in relevant metrics is a challenging task, which requires significant networks and systems expertise.

A second challenge comes from the fact that, in some cases, collecting data from only one equipment is not enough to fully comprehend a phenomenon. For example, localization data from a user device might need to be correlated with data collected by a network probe, raising interesting methodological questions.

The best way to describe my interest in this field is probably the Platonic allegory of the cave. In this cave, where a wood fire was crackling, Plato saw only the shadow of the men wandering therein. And the shadows projected on the walls of the cave were huge, nearly giving the impression they were those of giants. Metrology confronts us with this Platonic problem, as it only shows us the effects of all the mechanics of the network, while letting us imagine what might have created this outcome.

5.3.3 Asymmetric networks

Finally, a third subject I would like to further explore is the study of networks with asymmetric UL and DL. Most networking solutions are designed with the assumption of a symmetric communication link between hosts. Of course, asymmetric protocols, based on a master-slave approach, are classical in networking. But these asymmetric protocols still consider symmetric links, where slaves can instantly acknowledge receptions or announce the master of their intentions.

The last few years saw the emergence of networks with asymmetric links. Two examples come to my mind: low-power wide area networks (LPWANs), where the downlink, when present, is highly limited, and visible light communications (VLC), where a lighting bulb can be used to transmit data similarly to a base station, but the uplink is generally absent.

Current work in this field is more focused on very specific properties of these technologies, such as adapting carrier sense mechanisms to LPWAN [74] or proposing new modulations for VLC [34]. However, it is my belief that these networks do not need minor tweaks and adjustments of classical network protocols, but a completely new architectural design, based on a theory of asymmetric networks.

References

[1] 3GPP TR 38.874 V0.1.0 (2018). Study on Integrated Access and Backhaul. [Release 15].

[2] Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). Social fMRI: Investigating and Shaping Social Mechanisms in the Real World. *Elsevier Pervasive and Mobile Computing*, 7(6):643–659.

[3] Aijaz, A., Dohler, M., Aghvami, A. H., Friderikos, V., and Frodigh, M. (2017). Realizing the Tactile Internet: Haptic Communications over Next Generation 5G Cellular Networks. *IEEE Wireless Communications*, 24(2):82–89.

[4] Almeida, S., Queijo, J., and Correia, L. M. (1999). Spatial and Temporal Traffic Distribution Models for GSM. In *Proceedings IEEE VTC Fall 1999 - Vehicular Technology Conference*, Amsterdam, Netherlands.

[5] Andrews, J. G., Baccelli, F., and Ganti, R. K. (2011). A Tractable Approach to Coverage and Rate in Cellular Networks. *IEEE Transactions on Communications*, 59(11):3122–3134.

[6] Aoun, B., Boutaba, R., Iraqi, Y., and Kenward, G. (2006). Gateway Placement Optimization in Wireless Mesh Networks with QoS Constraints. *IEEE Journal on Selected Areas in Communications*, 24(11):2127–2136.

[7] Balachandran, A., Aggarwal, V., Halepovic, E., Pang, J., Seshan, S., Venkataraman, S., and Yan, H. (2014). Modeling Web Quality-of-Experience on Cellular Networks. In *Proceedings ACM MobiCom 2014 - International Conference on Mobile Computing and Networking*, Maui, HI, USA.

[8] Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., and Lepri, B. (2015). A Multi-source Dataset of Urban Life in the City of Milan and the Province of Trentino. *Nature Scientific Data*, 2(150055):1–15.

[9] Bhaumik, S., Chandrabose, S. P., Jataprolu, M. K., Kumar, G., Muralidhar, A., Polakos, P., Srinivasan, V., and Woo, T. (2012). CloudIQ: A Framework for Processing Base Stations in a Data Center. In *Proceedings ACM MobiCom 2012 - International Conference on Mobile Computing and Networking*, Istanbul, Turkey.

[10] Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., and Ziemlicki, C. (2012). Data for Development: the D4D Challenge on Mobile Phone Data.

[11] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[12] Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188.

[13] Bui, N., Cesana, M., Hosseini, S. A., Liao, Q., Malanchini, I., and Widmer, J. (2017). A Survey of Anticipatory Mobile Networking: Context-Based Classification, Prediction Methodologies, and Optimization Techniques. *IEEE Communications Surveys & Tutorials*, 19(3):1790–1821.

[14] Cai, L. and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14(2):1–10.

[15] Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2011). Estimating Origin-Destination Flows using Mobile Phone Location Data. *IEEE Pervasive Computing*, 10(4):36–44.

[16] Ceselli, A., Fiore, M., Premoli, M., and Secci, S. (2019). Optimized Assignment Patterns in Mobile Edge Cloud Networks. *Computers & Operations Research*, 106:246–259.

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

[18] Cici, B., Gjoka, M., Markopoulou, A., and Butts, C. T. (2015). On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology. In *Proceedings ACM MobiHoc 2015 - International Symposium on Mobile Ad Hoc Networking and Computing*, Hangzhou, China.

[19] Cisco (2019). Visual Networking Index - Mobile Forecast Highlights, 2017-2022.

[20] Coscia, M., Rinzivillo, S., Giannotti, F., and Pedreschi, D. (2012). Optimal Spatial Resolution for the Analysis of Human Mobility. In *Proceedings IEEE/ACM ASONAM 2012 - International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, Turkey.

[21] de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the Crowd: The Privacy Bounds of Human Mobility. *Nature Scientific Reports*, 3(1376).

[22] de Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. (2014). D4D-Senegal: The Second Mobile Phone Data for Development Challenge.

[23] De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., and Lepri, B. (2016). The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective. In *Proceedings ACM WWW 2016 - International Conference on World Wide Web*, Montreal, QC, Canada.

[24] Dhillon, H. S. and Caire, G. (2015). Wireless Backhaul Networks: Capacity Bound, Scalability Analysis and Design Guidelines. *IEEE Transactions on Wireless Communications*, 14(11):6043–6056.

[25] Domga Komguem, R., Stanica, R., Tchuente, M., and Valois, F. (2014). WARIM: Wireless Sensor Network Architecture for a Reliable Intersection Monitoring. In *Proceedings IEEE ITSC 2014 - International Conference on Intelligent Transportation Systems*, Qingdao, China.

[26] Domga Komguem, R., Stanica, R., Tchuente, M., and Valois, F. (2017). Node Ranking in Wireless Sensor Networks with Linear Topology. In *Proceedings IFIP WD 2017 - Wireless Days Symposium*, Porto, Portugal.

[27] Domga Komguem, R., Stanica, R., Tchuente, M., and Valois, F. (2019). Sensor Deployment in Wireless Sensor Networks with Linear Topology using Virtual Node Concept. *Springer Wireless Networks*, 26(164):1–16.

[28] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings TCC 2006 - International Conference on Theory of Cryptography*, New York, NY, USA.

[29] Elshaer, H., Boccardi, F., Dohler, M., and Irmer, R. (2015). Load & Backhaul Aware Decoupled Downlink/Uplink Access in 5G Systems. In *Proceedings IEEE ICC 2015 - International Conference on Communications*, London, England, UK.

[30] Finamore, A., Mellia, M., Gilani, Z., Papagiannaki, K., Erramilli, V., and Grunenberger, Y. (2013). Is There a Case for Mobile Phone Content Pre-Staging? In *Proceedings ACM CoNEXT 2013 - ACM Conference on Emerging Networking Experiments and Technologies*, Santa Barbara, CA, USA.

[31] Fooladivanda, D. and Rosenberg, C. (2013). Joint Resource Allocation and User Association for Heterogeneous Wireless Cellular Networks. *IEEE Transactions on Wireless Communications*, 12(1):248–257.

[32] Fortunato, S. (2010). Community Detection in Graphs. *Physics Reports*, 486(3):75–174.

[33] Furletti, B., Gabrielli, L., Renso, C., and Rinzivillo, S. (2012). Identifying Users Profiles from Mobile Calls Habits. In *Proceedings ACM UrbComp 2012 - ACM SIGKDD International Workshop on Urban Computing*, Beijing, China.

[34] Gancarz, J., Elgala, H., and Little, T. D. (2013). Impact of Lighting Requirements on VLC Systems. *IEEE Communications Magazine*, 51(12):34–41.

[35] Gavric, K., Brdar, S., Culibrk, D., and Crnojevic, V. (2013). Linking the Human Mobility and Connectivity Patterns with Spatial HIV Distribution. In *Proceedings NetMob 2013 - International Conference on the Analysis of Mobile Phone Datasets*, Boston, MA, USA.

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

[37] Gomez, C., Catalan, M., Figueras, X., Paradells, J., and Calveras, A. (2006). Impact of Handover between UMTS and GPRS on TCP/IP: An Empirical Approach. In *Proceedings IEEE VTC Fall 2006 - Vehicular Technology Conference*, Montreal, QC, Canada.

[38] Gomez, K., Goratti, L., Rasheed, T., and Reynaud, L. (2014). Enabling Disaster-resilient 4G Mobile Communication Networks. *IEEE Communications Magazine*, 52(12):66–73.

[39] Gomez-Miguelez, I., Garcia-Saavedra, A., Sutton, P. D., Serrano, P., Cano, C., and Leith, D. J. (2016). srsLTE: An Open-source Platform for LTE Evolution and Experimentation. In *Proceedings ACM WiNTECH 2016 - International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*, New York, NY, USA.

[40] Gramaglia, M. and Fiore, M. (2015). Hiding Mobile Traffic Fingerprints with GLOVE. In *Proceedings ACM CoNEXT 2015 - International Conference on Emerging Networking Experiments and Technologies*, Heidelberg, Germany.

[41] Haklay, M. and Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 7(4):12–18.

[42] Hawilo, H., Shami, A., Mirahmadi, M., and Asal, R. (2014). NFV: State of the Art, Challenges, and Implementation in Next Generation Mobile Networks (vEPC). *IEEE Network*, 28(6):18–26.

[43] Hinneburg, A. and Keim, D. A. (1999). Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. In *Proceedings ACM VLDB 1999 - International Conference on Very Large Databases*, Edinburgh, Scotland, UK.

[44] Horn, C., Klampfl, S., Cik, M., and Reiter, T. (2014). Detecting Outliers in Cell Phone Data: Correcting Trajectories to Improve Traffic Modeling. *Transportation Research Record: Journal of the Transportation Reserach Board*, 2405(1):49–56.

[45] I, C.-L., Huang, J., Duan, R., Cui, C., Jiang, J., and Li, L. (2006). Recent Progress on C-RAN Centralization and Cloudification. *IEEE Access*, 2:1030–1039.

[46] Kondor, D., Hashemian, B., de Montjoye, Y.-A., and Ratti, C. (2018). Towards Matching User Mobility Traces in Large-scale Datasets. *IEEE Transactions on Big Data*, 1(1):1–12.

[47] Lancichinetti, A. and Fortunato, S. (2009). Community Detection Algorithms: A Comparative Analysis. *Physics Review E*, 80(5):1–12.

[48] Li, N., Li, T., and Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proceedings ICDE 2007 - International Conference on Data Engineering*, Istanbul, Turkey.

[49] Li, R., Zhang, C., Patras, P., Stanica, R., and Valois, F. (2018). Learning Driven Mobility Control of Airborne Base Stations in Emergency Network. In *Proceedings IFIP WAIN 2018 - International Workshop on Artificial Intelligence in Networks*, Toulouse, France.

[50] Liu, C., Sundaresan, K., Jiang, M., Rangarajan, S., and Chang, G.-K. (2013). The Case for Re-Configurable Backhaul in Cloud-RAN based Small Cell Networks. In *Proceedings IEEE Infocom 2013 - International Conference on Computer Communications*, Turin, Italy.

[51] Liu, D., Wang, L., Chen, Y., Elkashlan, M., Wong, K.-K., Schober, R., and Hanzo, L. (2016). User Association in 5G Networks: A Survey and an Outlook. *IEEE Communications Surveys & Tutorials*, 18(2):1018–1044.

[52] Lyu, J., Zeng, Y., Zhang, R., and Lim, T. J. (2017). Placement Optimization of UAV-Mounted Mobile Base Stations. *IEEE Communications Letters*, 21(3):604–607.

[53] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). L-Diversity: Privacy Beyond k-Anonymity. In *Proceedings ICDE 2006 - International Conference on Data Engineering*, Atlanta, GA, USA.

[54] Milligan, G. W. and Cooper, M. C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50(2):159–179.

[55] Moradi, M., Sundaresan, K., Chai, E., Rangarajan, S., and Mao, Z. M. (2018). SkyCore: Moving Core to the Edge for Untethered and Reliable UAV-based LTE Networks. In *Proceedings ACM MobiCom 2018 - International Conference on Mobile Computind and Networking*, New Delhi, India.

[56] Naboulsi, D. and Fiore, M. (2013). On the Instantaneous Topology of a Large-scale Urban Vehicular Network: The Cologne Case. In *Proceedings ACM MobiHoc 2013 - ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Bangalore, India.

[57] Naboulsi, D., Fiore, M., and Stanica, R. (2013). Human Mobility Flows in the City of Abidjan. In *Proceedings NetMob 2013 - International Conference on the Analysis of Mobile Phone Datasets*, Boston, MA, USA.

[58] Nikaein, N., Marina, M. K., Manickam, S., Dawson, A., Knopp, R., and Bonnet, C. (2014). OpenAirInterface: A Flexible Platform for 5G Research. *ACM SIGCOMM Computer Communication Review*, 44(5):33–38.

[59] Official Journal of the European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).

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

[61] Oueis, J., Conan, V., Lavaux, D., Rivano, H., Stanica, R., and Valois, F. (2019). Core Network Function Placement in Self-Deployable Mobile Networks. *Computer Communications*, 133(1):12–23.

[62] Peng, C., Lee, S.-B., Lu, S., Luo, H., and Li, H. (2011). Traffic-Driven Power Saving in Operational 3G Cellular Networks. In *Proceedings ACM MobiCom 2011 - International Conference on Mobile Computind and Networking*, Las Vegas, NV, USA.

[63] Piorkowski, M., Sarafijanovic-Djukic, N., and Grossglauser, M. (2009). A Parsimonious Model of Mobile Partitioned Networks with Clustering. In *Proceedings COMSNETS 2009 - International Conference on Communication Systems and Networks*, Bangalore, India.

[64] Qazi, Z. A., Lee, J., Jin, T., Bellala, G., Arndt, M., and Noubir, G. (2013). Application-Awareness in SDN. In *Proceedings ACM Sigcomm 2013 - International Conference on Data Communications*, Hong Kong, China.

[65] QGIS Geographic Information System (2019). Open Source Geospatial Foundation Project. https://qgis.org/. [Online; accessed 30-August-2019].

[66] Shafiq, M. Z., Erman, J., Ji, L., Liu, A. X., Pang, J., and Wang, J. (2014). Understanding the Impact of Network Dynamics on Mobile Video User Engagement. *ACM SIGMETRICS Performance Evaluation Review*, 42(1):367–379.

[67] Shafiq, M. Z., Ji, L., Liu, A. X., Pang, J., Venkataraman, S., and Wang, J. (2013a). A First Look at Cellular Network Performance during Crowded Events. In *Proceedings ACM SIGMETRICS 2013 - International Conference on Measurement and Modeling of Computer Systems*, Pittsburgh, PA, USA.

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

[69] Sheltami, T. R., Shakshuki, E. M., and Mouftah, H. T. (2009). Performance Evaluation of TelosB Sensor Network. In *Proceedings MoMM 2009 - International Conference on Advances in Mobile Computing and Multimedia*, Kuala Lumpur, Malaysia.

[70] Stoica, A., Smoreda, Z., Prieur, C., and Guillaume, J.-L. (2010). Age, Gender and Communication Networks. In *Proceedings NetMob 2010 - International Conference on the Analysis of Mobile Phone Datasets*, Boston, MA, USA.

[71] Sweeney, L. (2002). k-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

[72] Tang, J., Tay, W. P., Quek, T. Q. S., and Liang, B. (2017). System Cost Minimization in Cloud RAN with Limited Fronthaul Capacity. *IEEE Transactions on Wireless Communications*, 16(5):3371–3384.

[73] Tene, O. and Polonetsky, J. (2013). Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property*, 11(5):240–273.

[74] To, T.-H. and Duda, A. (2018). Simulation of LoRa in NS-3: Improving LoRa Performance with CSMA. In *Proceedings IEEE ICC 2018 - IEEE International Conference on Communications*, Kansas City, MO, USA.

[75] Trestian, I., Ranjan, S., Kuzmanovic, A., and Nucci, A. (2009). Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network. In *Proceedings ACM IMC 2009 - International Conference on Internet Measurement*, Chicago, IL, USA.

[76] Uppoor, S., Trullols-Cruces, O., Fiore, M., and Barcelo-Ordinas, J. M. (2013). Generation and Analysis of a Large-scale Urban Vehicular Mobility Dataset. *IEEE Transactions on Mobile Computing*, 13(5):1061–1075.

[77] Verleysen, M. and Francois, D. (2005). The Curse of Dimensionality in Data Mining and Time Series Prediction. In *Proceedings IWANN 2005 - International Work-Conference on Artificial Neural Networks*, Barcelona, Spain.

[78] Vieira, M. R., Frias-Martinez, V., Oliver, N., and Frias-Martinez, E. (2010). Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics. In *Proceedings IEEE SocialCom 2010 - International Conference on Social Computing*, Minneapolis, MN, USA.

[79] Wang, K., Zhao, M., and Zhou, W. (2014). Graph-based Dynamic Frequency Reuse in Cloud-RAN. In *Proceedings IEEE WCNC 2014 - Wireless Communications and Networking Conference*, Istanbul, Turkey.

[80] Xu, R. and Wunsch, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.

[81] Yuan, J., Zheng, Y., Xie, X., and Sun, G. (2011). Driving with Knowledge from the Physical World. In *Proceedings ACM KDD 2009 - SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA.

[82] Zang, H. and Bolot, J. (2011). Anonymization of Location Data Does Not Work: A Large-scale Measurement Study. In *Proceedings ACM MobiCom 2011 - International Conference on Mobile Computing and Networking*, Las Vegas, NV, USA.

[83] Zang, H. and Bolot, J. C. (2007). Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks. In *Proceedings ACM MobiCom 2007 - International Conference on Mobile Computing and Networking*, Montreal, QC, Canada.

[84] Zheng, Y., Liu, L., Wang, L., and Xie, X. (2008). Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. In *Proceedings ACM WWW 2008 - International Conference on World Wide Web*, Beijing, China.

[85] Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *Proceedings ACM WWW 2009 - International Conference on World Wide Web*, Madrid, Spain.

[86] Zhu, Z., Gupta, P., Wang, Q., Kalyanaraman, S., Lin, Y., Franke, H., and Sarangi, S. (2011). Virtual Base Station Pool: Towards a Wireless Network Cloud for Radio Access Networks. In *Proceedings ACM CF 2011 - International Conference on Computing Frontiers*, Ischia, Italy.