Characterization of mobile network services to assess the impact of network slicing in a nationwide scenario

María Cristina Márquez Colás

Thesis deposited in partial fulfillment of the requirements for the degree of Doctor of Philosophy (PhD) in

Telematics Engineering

Universidad Carlos III de Madrid

Advisors:

Albert Banchs
Marco Gramaglia

September 2020

Me: *I don't see myself with a PhD.*
(Gets the minimum requirements for a PhD in the first year.)
Cristina Marquez (2015)

Friends and family: *Do you ever do anything the easy way?*
Me: *And risk disappointing you?)*
Cristina Marquez (using Lara Croft words).

Education is what is left after all that has been learnt is forgotten.
James Bryant Conant - Diary as a freshman at Harvard College (1910)

Everything is connected to Everything else...
Barry Commoner - biologist (1917)

Just remember to have fun.
Cristina Marquez, pharaprasing Quistis advice (2020)

Acknowledgements

Another year passes by, and another thesis is done. Hopefully, this will be the last one, as I could not only achieve the highest Education degree in Telematics Engineering, but also some of my bucket list wishes with additional non-expected good times.

I would like to summarize this PhD experience with some of the best Lara Croft's quotes:

- *I'd finally set out to make my mark; to find adventure. But instead adventure found me.* - PhD offer.
- *The world is full of unanswered questions, beyond all limits or reason... the answers await.* - PhD starts.
- *A famous explorer once said, that the extraordinary is in what we do, not who we are.* - First publications accepted.
- *Everything lost is meant to be found.* - Dream about going to MIT.
- *I make my own luck.* - Selected as Next Generation Of Internet Explorer.
- *In our darkest moments, when life flashes before us, we find something; Something that keeps us going. Something that pushes us.* - Related to the COVID-19 pandemic experience from the US.
- *A scar means I survived.* - Deposit and defense of the thesis.
- *The fate of humanity is now in (y)our hands.* - New chapter in my life, and in my mentees one.

Thanks to my advisors and colleagues from Spain and US for all the knowledge and hard-work. This path would have not been the same without you.

Over these three years, I could not only learn a lot and challenge myself while maintaining my connections with my loved ones, but also travel and meet new people. A special thank you goes to my family for supporting me in every adventure, despite how far I was going.

Similarly, my supporters and always available helping hands deserve a huge thank you. You helped me to grow professionally and personally, staying by my side in the good and the bad times, even at the distance, over all these years.

Also, I want to appreciate some other words that made me stronger and always keep my force of will : "you will do anything you want in life", "your CV scared me, you were too good", "you are the next woman that would influence ICT / STEM".

Last but not least, to all of you, here is part of my work. I hope it would be used for good in the next 5th Generation of Networks, Games, Smart cities and Society as an end point.

Published and submitted content

The thesis is based on the following published papers:

[1] **C. Marquez**, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in Proceedings of *the 13th International Conference on Emerging Networking EXperiments and Technologies (ACM CoNEXT 2017)*, Incheon/Seoul, Republic of South Korea, Dec. 2017, p.180–186. [Online]. Available: https://doi.org/10.1145/3143361.3143369

- This work is partially included in this thesis and its content is reported in Chapter 1, 2, and 3.
- The author's role in this work is focused on analyzing and classifying one week time-series of real mobile traffic data, normalilzing and computing statistics of the traffic volume, extracting mobile service distributions, developing the algorithm to detect activity peaks, correlating spatio-temporal per-user traffic and drawing up conclusions about the mobile services (i.e., applications).

[2] **C. Marquez**, M. Gramaglia, M. Fiore, A. Banchs, and Z. Smoreda, "Identifying Common Periodicities in Mobile Service Demands with Spectral Analysis," in Proceedings of *the 18th Mediterranean Communication and Computer Networking Conference (IEEE MedComNet 2020)*, Arona, Italy, Jun. 2020, pp.1-8. [Online]. Available: https://doi.org/10.1109/MedComNet49392.2020.9191477

- This work is partially included in this thesis and its content is reported in Chapter 1, and 4.
- The author's role in this work is focused on the design, implementation and experimentation of a two-step spectral algorithm to identify and classify the most retaining frequency components of three-months time-series of real mobile applications with regarding of the concepts proposed in the paper, aiming to explain their similarities and exploiting them for applications in network planning and resource management.

[3] **C. Marquez**, M. Gramaglia, M. Fiore, A. Banchs, and Costa-Pérez, "Resource Sharing Efficiency in Network Slicing," in *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 909–923, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8737701

- This work is partially included in this thesis and its content is reported in Chapter 6.
- The author's role in this work is focused on design, implementation and experimentation of an innovative data-driven resource allocation technique that guarantees a percentage of the traffic demand and allows the operator to overbook the network resources in distinct scenarios that could be compared with the reference use cases in [4] with regarding of the concepts and hierarchical network structure proposed in the paper.

[4] **C. Marquez**, M. Gramaglia, M. Fiore, A. Banchs, and Costa-Perez, "How Should I Slice My Network? A Multi-Service Empirical Evaluation of Resource Sharing Efficiency," in Proceedings of *the 24th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2018)*, New Delhi, India, 2018, p. 191–206. [Online]. Available: https://dl.acm.org/doi/10.1145/3241539.3241567

- This work is partially included in this thesis and its content is reported in Chapter 2, and 6.
- The author's role in this work is focused on the data-driven design, implementation and experimentation of static and dynamic resource allocation techniques of real mobile services data in network slices, as well as the load balancing configuration and association of antenna sites as defined in the paper's case studies and scenarios, while guaranteeing the slice specification.

[-] **C. Marquez**, J. Á. Fernández-Rodrigues, M. Gramaglia, M. Fiore, A. Banchs, and Cezary Ziemlicki, working paper under preparation.

- This working paper is partially included in this thesis and its content is reported in Chapter 5.
- The author's role in this work is focused on the experimentation of a new service clusterization approach using Wavelet Transformation (WT), explaining how it relates to the previous techniques that considered a spatio-temporal and spectral methodology. In addition, the author draws up conclusions about mobile services and the advantages of following this approach in distinct scenarios for network planning and resource management purposes.

Abstract

Several business are nowadays becoming more and more aware of the potential that lies beneath Big Data. From social media 'titans' and healthcare companies, to the mobile industry that propelled them, 5th Generation mobile network (5G) will be a fundamental factor that will drive our new reality. Current mobile service usage has been explored for data-driven organizational growth in the touristic sector, as it allows forecasting hotel occupancy rates or targeting customersHowever, mobile data is continuously increasing, and therefore it is enormously important to analyze such data for networking purposes.

Previous Ericsson Mobility Reports claimed that by the end of 2022 the total monthly traffic associated with mobile devices would be 77 exabytes (EB), representing 20% of the total Internet Protocol (IP) traffic around the world. They also declare that 50 EB/month would come from 2nd Generation mobile network (2G), 3rd Generation mobile network (3G), and 4th Generation mobile network (4G) devices. In terms of 5G subscriptions, it is expected to reach up to 2.8 billion subscriptions globally by the end of 2025, accounting for about 30% of total mobile subscriptions. Experts stake out that by the year 2020, 1.7 megabytes of data will be generated every second for every person on the planet, forcing the network to evolve and adapt to challenging new demands.

Network providers not only deal with the deployment of the required resources to support this growth, but also the potential newcomers into the business, and the stringent conditions of the distinct services to be provided. One of the features proposed to face this dilemma is using the network slicing technique. It allows to transform and orchestrate a 5G network by creating multiple logical instances (*i.e.*, slices) on top of it, while Big Data would provide the specifications of the services' traffic dynamics to be served. In this way, operators achieve the best allocation of resources.

This thesis contributes to the ongoing Network Slicing research, assessing a nationwide scenario. Our results show mobile traffic similarities and differences across time, space, and frequency domains, whereas we intend for distinct service clusterizations that would enhance the network efficiency in terms of resource management. For instance, we show that benefits are achieved when considering the top 10 consuming Network Slice (NS). In addition, we could observe mobile service similarities in the spatial domain, while the spectral and time domains open the door for wavelets uncertainty, where we point future directions to address this research branch. Moreover, we propose two data-driven algorithms that shed light on the trade-off between complexity and multiplexing efficiency derived from the network slice specifications, both exhibiting promising performances (*e.g.*, leading to a new architecture for traffic balancing in the cloud and edge clusters, with 60% and 400% gain in efficiency respectively and 1/3 of dedicated resources).

Keywords: Big Data, Network slicing, Resource Management, Network Efficiency, Mobile Networks, Slice Orchestration, NFV.

Contents

List of Figures

List of Tables

List of Acronyms

2G 2nd Generation mobile network.

3G 3rd Generation mobile network.

3GPP 3rd Generation Partnership Project.

4G 4th Generation mobile network.

5G 5th Generation mobile network.

AI Artificial Intelligence.

BBUs Base Band Units.

C-RAN Cloud/Centralized Radio Access Network.

CAPEX CAPital EXpenditure.

CDF Cumulative Density Function.

CDRs Call Detail Records.

CNIL Commission Nationale de l'Informatique et des Libertés.

CORE COmputing Research and Education.

CPU Central Processing Unit.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DFT Discrete Fourier Transform.

DFTs Discrete Fourier Transform.

DPI Deep Packet Inspection.

DPO Data Privacy Officer.

EB exabytes.

EC European Commission.

eMBB Enhanced Mobile Broadband.

EPS Evolved Packet System.

EUTRAN Evolved Universal Mobile Telecommunications System Terrestrial Radio Access Network.

FFT Fast Fourier Transform.

FFTs Fast Fourier Transforms.

GDPR General Data Protection Regulation.

GGSN Gateway GPRS Support Node.

GTP-C GPRS Tunneling Protocol Control.

GTP-U GPRS Tunneling Protocol User.

IoT Internet of Things.

IP Internet Protocol.

ITU International Telecommunication Union.

JCR Journal Citation Reports.

KaFFPa Karlsruhe Fast Flow Partitioner.

KPI Key Performance Indicator.

KPIs Key Performance Indicators.

LTE Long Term Evolution.

MAC Medium Access Control.

MANO Management and Orchestration.

MEC Mobile Edge Computing.

ML Machine Learning.

MMS Multimedia Messaging Service.

mMTC massive Machine Type Communication.

NFs Network Functions.

NFV Network Function Virtualization.

NFVI Network Functions Virtualization Infrastructure.

NGMN Next Generation Mobile Networks.

NS Network Slice.

OPEX OPerating EXpense.

P-GW Packet Data Network (PDN) Gateway.

P2P Peer-to-peer.

PCA Principal Component Analysis.

PDF Probability Density Function.

PDP Packet Data Protocol.

PhD Doctor of Philosophy.

QoS Quality of Service.

RA/TA Routing/Tracking Areas.

RAN Radio Access Network.

SDM-C Software-Defined Mobile network Control.

SDM-X Software-Defined Mobile network Coordinator.

SDN Software Defined Networking.

SLA Service Level Agreement.

SLAs Service Level Agreements.

TGV high-speed "Train à Grande Vitesse" train.

ULI User Location Information.

URLLC Ultra Reliable Low Latency Communication.

UTRAN Universal Mobile Telecommunications System Terrestrial Radio Access Network.

VLAN Virtual Local Area Network.

VM Virtual Machine.

VMs Virtual Machines.

VNF Virtual Network Function.

VNFs Virtual Network Functions.

VoIP Voice over IP.

Wi-Fi Wireless Fidelity.

WT Wavelet Transformation.

1 Introduction

Contents

It is worldwide agreed that current mobile networks will have to experience a huge revolution to cover all the needs of the users and vertical customers, as well as the management of their generated data. This revolution is known as 5G, and in order to boost that leap, there are several techniques to explore and areas to improve.

In this first chapter of the thesis, we introduce the motivation underneath the characterization of mobile network services elaborated in the remainder of this work (see Section 1.1), as well as the landscape that the new mobile generation pose for current network technologies (see Section 1.2). Next, we present the goals and contributions of the developed work in Section 1.3. To end up, we outline the structure of the document per se (see Section 1.4). Hence, this chapter is divided in four sections in order to address each one of these topics.

1.1 Motivation

As the mobile Internet traffic grows along with the quantity and diversity of offered services, it becomes increasingly important to understand the demands generated by them. Indeed, characterizing the traffic dynamics associated to different mobile services is of paramount importance in order to properly dimension and orchestrate the mobile network, and also offers an opportunity to unravel broader societal behaviors in general. In social sciences, understanding the dynamics of the demands for mobile services helps drawing causal links between land use and the way mobile applications are used [5], highlighting cultural factors in app adoption [6], helping governments to take informed public health actions [7], or even detecting psychiatric disorder states at scale [8]. From an engineering and technology viewpoint, the knowledge of large-scale

traffic volumes generated by each mobile service enables a more efficient dimensioning and management of the communication infrastructure [4], the optimized caching of applications data at mobile devices [9], the interplay between the digital world space and the physical one (*e.g.*, urban development [10] or planning [5,11]), or the improved planning of urban transport systems based on app user flows [12].

While disentangling mobile service demand patterns is a challenging exercise, it is necessary for a proper allocation of resources and shaping networks accordingly, which is a fundamental operation for 5G networks.The many and varied apps running on mobile devices entail strongly heterogeneous dynamics (*e.g.*, high data rates, sub-ms delays, extensive coverage) [13, 14] over a space that is high-dimensional along both its temporal (where measurement data can encompass long periods of months with a fine granularity of minutes) and geographical (with traffic information concurrently recorded at hundreds of locations within, *e.g.*, a single metropolitan area) facets. In addition, measurements are often noisy, due to inherent randomness in user access to apps [15], oscillations in device associations to the radio access infrastructure caused by signal strength fluctuations, load balancing policies [16], or positioning accuracy limitations of the mobile network technology [17].

Therefore, network operators and tenants need to cope with all these issues and performance metrics in a new infrastructure with respect to Long Term Evolution (LTE) [18, 19]. On the one hand, one of the new enablers is Network Function Virtualization (NFV), a general reference framework for the network architecture. Its aim is to virtualize different elements of the network, motivated by the necessity of more flexible architectures. On the other hand, Software Defined Networking (SDN) is a paradigm that allows to decouple the Radio Access Network (RAN) control functionality from the data plane. We can see the difference between the SDN approach and the classical one in Figure 1.1. Hence, SDN simplifies the network deployment and operation along with reducing the total cost of managing networks by means of network slices and NFV. For instance, an effective orchestration of network slices (virtual networks with dedicated resources allocated and customized Quality of Service (QoS) guarantees and network Key Performance Indicators (KPIs)) could be built on the spatial complementarity of the demands for the different services [20].

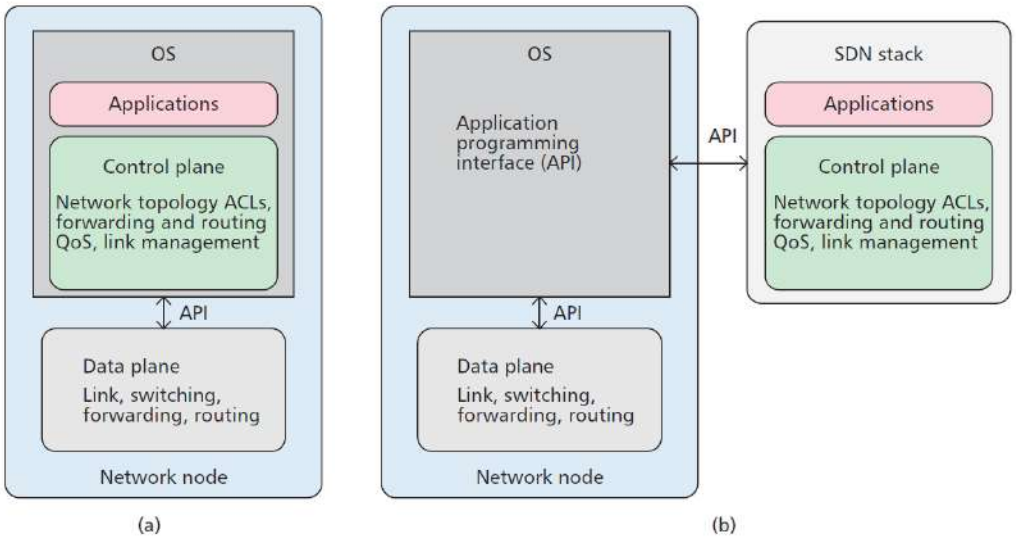


Figure 1.1. Traditional network (a) vs SDN network (b) [21]

To the best of our knowledge, only a couple of works analyzed a large set of heterogeneous individual mobile services to date. They focus on quantifying the statistical diversity of apps in terms of network-level performance indicators [22], and forecasting future demands for each app separately [23], respectively. As a result, most previous works in the area are concerned with coarse-grained service categories or rely on relatively small-sized datasets, providing limited knowledge about the specificities of individual services and their differences.

In this thesis, we contribute to fill the gap above, by analyzing the usage of a selection of mobile services at a national scale and in two residential areas with regards of 5G networks. Our study investigates the traffic behavior of specific services across time (*i.e.*, temporal usage patterns), space (*i.e.*, at different locations), and spectral (*i.e.*, frequency components) domains. In doing so, we offer a global overview of the traffic dynamics for specific applications in a large-scale operational cellular network.

Furthermore, the results and insights gained from our analysis may find applications in different areas. In future-generation mobile networks, the understanding of when, where and how different mobile services are consumed will be essential to dynamically tailor resources to the actual fluctuations of the subscribers' activity. Indeed, many novel architectural paradigms aim at enabling the dynamic management of system resources, across multiple network functions at the network edge or core [21, 24, 25].

1.2 Landscape

In the section above, we described the concept of a network slice, but we did not provide details about how they run. As it can be seen in Figure 1.2, a Network Slice Instance operates on top of a shared infrastructure, which is composed of generic and dedicated hardware resources.

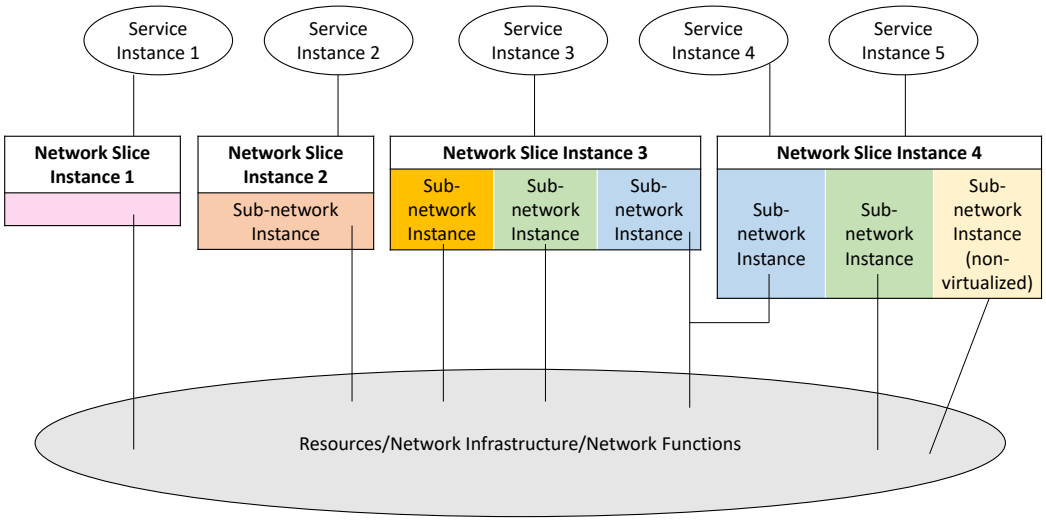


Figure 1.2. Network slicing conceptual outline [26]

We also introduce the NFV framework. However, the complexity of such a paradigm is a challenge. As depicted in Figure 1.3, we can divide it into the following blocks:

- Network Functions Virtualization Infrastructure (NFVI), which contains both the hardware that allocates the Virtual Machines (VMs) and the program that allows to virtualize resources.
- Virtual Network Function (VNF), which uses the VMs from the NFVI block. Moreover, it adds the required software upon the VNFs.
- Management and Orchestration (MANO), which is a disconnected block in the architecture that interacts with both NFVI and VNF blocks. Here, all the resource management base is delegated (including the space reservation, creation, and deletion of the VMs).

Furthermore, there are additional elements to manage and control multiple slices. For instance, the Software-Defined Mobile network Coordinator (SDM-X) is required in order to ensure high resource efficiency and individual Service Level Agreements (SLAs) when many network slices share Network Functions (NFs). Another important entity is the Software-Defined Mobile network Control (SDM-C), on which network slices rely. It includes not only virtual NFs, but also physical ones [27].

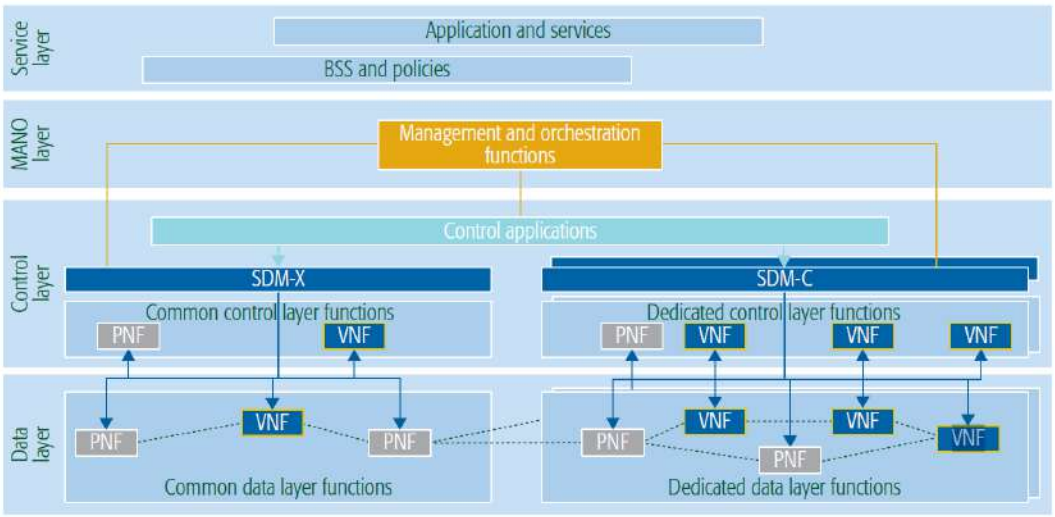


Figure 1.3. Schema of 5G network layers [27]

Considering that all the functions should run over the RAN, the protocol stack needs a re-design. Some implementations showed a low performance when computational outages occur [28], but recent studies [29] provide solutions with a better performance and temporarily limited computational resources in a virtualized RAN. Hence, the real problem comes with the nature of the services in the network.

According to the International Telecommunication Union (ITU) [30], the envisioned services instantiated in a specific network slice are: Enhanced Mobile Broadband (eMBB), massive Machine Type Communication (mMTC), and Ultra Reliable Low Latency Communication (URLLC). Since slices may serve heterogeneous services within the same infrastructure, different telecommunication services (arranged to a single slice) can be configured separately taking into account their unique characteristics. In addition, as slices run on a shared infrastructure, their cost-efficient customization is allowed via the cloudified network.

From a system standpoint, the technology needed to support the different types of slices is well understood or even already available. For instance, several cloud resource

orchestrators exist for both commercial and open-source telco-cloud platforms [31]; similarly, traditional tools for network management are insufficient for the amount of data flowing through the network and the difficulties in forecasting the behavior of a system involving many different slices. Therefore, a variety of solutions has been proposed for the dynamic allocation of resources across network slices [32]. However, multiplexing efficiency regarding resource management has not been deeply explored with real data and the parameters that may be needed for reconfiguration.

Thus, a network slicing framework should define novel algorithms to efficiently manage the infrastructure resources, sharing them between the different slices while guaranteeing that the requirements of each slice are met.

1.3 Objectives and contributions

As described above, it is required to understand service demands as well as the distinct requirements of each of them for a proper network slice definition. It is also imperative to analyze the trade-offs between network efficiency, complexity and customization of such services as close to reality as possible. In this way, operators will be able take informed decisions on how to orchestrate future networks.

This doctoral thesis is a first step to address these questions with a data-driven approach. Firstly, it includes an analysis of mobile services given their spatio-temporal and frequency characteristics. Secondly, it describes two algorithms that allow to efficiently manage the network resources by means of network slices. Our overall contributions are as follows:

1. **Spatio-temporal analysis.** Individual mobile services are inspectioned at a national scale, by studying data collected in a 3G/4G mobile network deployed over a major European country. Through correlation and clustering analyses, our study unveils a strong heterogeneity in the demand for different mobile services, both in time and space. In particular, we show that: *(i)* somehow surprisingly, almost all considered services exhibit quite different temporal usage patterns; *(ii)* in contrast to such temporal behavior, spatial patterns are fairly uniform across all services; *(iii)* when looking at usage patterns at different locations, the average traffic volume per user is dependent on the urbanization level, yet its temporal dynamics are not. Last but not least, our findings do not only have sociological implications, but are also relevant to the orchestration of network resources.

2. **Spectral analysis.** In the context of heterogeneity derived by the analysis above, in order to find common patterns across services, we hinge upon a spectral analysis framework, by computing Discrete Fourier Transform (DFTs) of the typical demands for tens of popular mobile services observed in an operational metropolitan-scale network. We filter, cluster, and analyse hundreds of frequency components, and identify a substantial set of regular patterns that are common across most service demands. We also unveil how several mobile services defy classification, and have instead highly distinguishing temporal dynamics.

3. **Hybrid time-frequency analysis.** As shown in the previous contributions, mobile traffic time-series could be transformed and decomposed in order to

find proper clusterization approaches. In order to exploit the conciseness of frequency analysis and the expressiveness of the temporal analysis, we followed an hybrid approach and we apply the wavelet transformation to extract relevant features both in the time and frequency domain of a collection of different mobile services. This approach allows to extract their information without losing the connection between these two dimensions. Then, we apply unsupervised clustering algorithms to the relevant features in order to classify the behaviour and be able to provide recommendations on how to allocate network resources for distinct clusters.

4. **First cost-efficient customization algorithm.** As the economic sustainability of future mobile networks will largely depend on the strong specialization of its offered services, network operators will need to provide added value to their tenants, by moving from the traditional *one-size-fits-all* strategy to a set of virtual end-to-end instances of a common physical infrastructure (*i.e.*, *network slices*). We provide a first empirical investigation of the trade-off between: *(i)* the need for fully dedicated resources to support *service customization*, and *(ii)* the dynamic resource sharing among services to increase *resource efficiency* and cost-effectiveness of the system in network slicing scenarios. Building on substantial measurement data collected in an operational mobile network *(i)* we quantify the efficiency gap introduced by non-reconfigurable allocation strategies of different kinds of resources, from radio access to the core of the network, and *(ii)* we quantify the advantages of their dynamic orchestration at different timescales. Our results provide insights on the achievable efficiency of network slicing architectures, their dimensioning, and their interplay with resource management algorithms.

5. **Waterfilling algorithm.** Implementing network slicing has significant consequences in terms of resource management, as we showed with the previous algorithm. In this case, we adopt a novel data-driven approach to quantify the efficiency of resource sharing in future sliced networks under overbooking conditions, and we compare it to the previous one. Service customization entails assigning to each slice fully dedicated resources, which may also be dynamically reassigned and overbooked in order to increase the cost-efficiency of the system. Building on metropolitan-scale real-world traffic measurements, we carry out an extensive parametric analysis that highlights how diverse performance guarantees, technological settings, and slice configurations impact the resource utilization at different levels of the infrastructure in presence of network slicing.

The result of such contributions were published in the following publications. The first contribution was published in *Proceedings of the 13th International Conference on Emerging Networking EXperiments and Technologies (ACM CoNEXT) 2017*, indexed in COmputing Research and Education (CORE) A ranking. The second one was recently published in *the 18th Mediterranean Communication and Computer Networking Conference (IEEE MedComNet) 2020*. The third contribution belongs to a working paper under preparation. The fourth one was published in Proceedings of *the 24th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2018)*, indexed in CORE A* ranking. Finally, the last contribution

was published in *IEEE Transactions on Network and Service Management (IEEE TNSM) 2019* as a *special issue on Novel Techniques in Big Data Analytics for Management*, indexed in Journal Citation Reports (JCR).

1.4 Outline of the thesis

This thesis is divided in 7 chapters, plus the references. The chosen structure facilitates the reading comprehension and the achievement of the distinct objectives according to the following order:

- **Chapter 1: Introduction**. It describes current trends, motivations, technologies and challenges of the research field of interest. The structure of the thesis is summarized.

- **Chapter 2: State of art**. This chapter addresses the global framework for network slicing, providing the expected 5G agenda. It continues with an explanation about the network slicing strategies for 5G networks, the related work of Data Analysis studies, as well as a section with relevant information about the guaranteed legal usage of users' data in this research. We finish this chapter adding a brief section on security .

- **Chapter 3: Spatio-temporal analysis**. It starts describing the data employed. It continues with an overview of the mobile services extracted. Finally, the temporal and spatial clustering results are presented and discussed.

- **Chapter 4: Spectral analysis**. This chapter introduces a new methodology to cluster mobile services. In order to do so, we describe a preprocessing of the data. Then, the most relevant components are filtered for clustering purposes. After that, the commonalities and outliers in mobile services are reviewed.

- **Chapter 5: Hybrid time-frequency analysis**. Firstly, the wavelet representation is described. Secondly, a new methodology is introduced to characterize mobile services. Thirdly, a clusterization is applied over the new characterization data. Finally, we examine the same methodology over a region where services are distributed over a number of clusters in a specific area.

- **Chapter 6: Data-driven resource management**. In this chapter, we explore two data-driven algorithms that explore the trade-off between service customization and resource efficiency in network slicing scenarios. Once these scenarios are defined, the metrics and slice specification are described. Next, we explore some settings that allow to understand the relation between several variables, such as number of slices or time window, for static and dynamic resource orchestrated configurations.

- **Chapter 7: Conclusions and future work**. It draws the most important conclusions of the aforementioned research and future lines of work.

2

State of the art

Contents

As mentioned in Chapter 1, it is now commonly agreed that current networks will experience a change of paradigm in order to achieve the goals set for 5G mobile networks. To this end, they will be built upon the network slicing concept, that encompasses with it many other network changes.

In the first section, this Chapter defines the distinct key concepts of the 5G agenda and the motivation for network slicing. In the second section we describe the considered network architecture to define distinct network slices. Next, since our data comes from a European Operator, we address the related work associated with mobile phone data and the possible legal issues associated to such data in our work. Finally, the last section explores briefly the network security concerns about this new digital era.

2.1 5G technology

Current trends in mobile networks point towards a strong diversification of services, which are characterized by increasingly heterogeneous KPI and QoS requirements. This tendency is driving the design of 5G networks that will eventually have to support, *e.g.*, the Internet of Things (IoT) with ultra-low rate communication from a massive number of devices, automotive and tactile applications with millisecond latencies, industrial communications with extreme reliability, and virtual/augmented reality services with very high data rates.

However, clear needs for tailored KPI and QoS requirements are already evident in today's mobile services, which encompass, *e.g.*, high-quality video streaming,

machine-type communication, low-latency mobile gaming, jointly with best effort traffic. Unfortunately, current mobile network architectures [33] lack the necessary flexibility to meet the extreme requirements imposed by such services.

This situation is pushing independent initiatives to address the problem. 3rd Generation Partnership Project (3GPP) has developed an Internet of Things (IoT)-specific Medium Access Control (MAC) that co-exists with the legacy general-purpose MAC layer [34]. Network deployments in industrial environments rely on proprietary architectures that ensure reliability levels not attainable with public mobile networks [35]. Google has started to deploy its own radio access infrastructure and proprietary transit networks to run its many services under hard QoS guarantees [36].

While the scope of the mentioned solutions is clearly limited, they do show the need for customized network support even with present-day traffic. They also substantiate the well-established vision that several network instances, each devoted to a specific set of services, have to co-exist in the same infrastructure in order to satisfy the KPI and QoS requirements of current and future mobile applications.

The agenda for 5G networks is to achieve this via some diversification of the physical network, and mainly via *network virtualization*, which evolves the traditional hardbox paradigm into a cloudified architecture where the once hardware-based network functions (*e.g.*, spectrum management, baseband processing, mobility management) are implemented as software VNFs running on a general-purpose *telco-cloud*. Network virtualization enables the deployment of multiple virtual instances of the complete network, named *network slices*.

Slices are then easily customized by tuning the functionality and location of VNFs. They thus create on top of the physical infrastructure a set of logical networks, each tailored to accommodate fine-tuned SLA reflecting the needs of different service providers.

2.2 Network slicing strategies

Even though we have already defined what a network slice is, and its context in future networks, the possible strategies are a topic we need to describe to fully understand our research. Since we have in mind the full picture of the new generation of networks described in Chapter 1, we can focus on network slicing.

Network slicing has profound implications on resource management. When instantiating a slice, an operator needs to allocate sufficient computational and communication resources to its VNFs. In some cases, these resources may be dedicated, becoming inaccessible to other slices [37]. Alternatively, smart assignment algorithms can be employed to dynamically allocate resources to slices based on the time-varying demands of tenants [38, 39]. This grants the flexibility to modify the share of resources assigned to each tenant, multiplexing logical slices into software or hardware assets while trying to abide by tenant requirements. However, such algorithms introduce additional complexity, and may in some cases hinder resource isolation, the corresponding guarantees to tenants, and/or the ability to deploy fully customized slices.

The above shows that there is an inherent trade-off among: *(i) service customization*, which favours the deployment of specialized slices with tailored functions for each service and, possibly, dedicated and guaranteed resources; *(ii)*

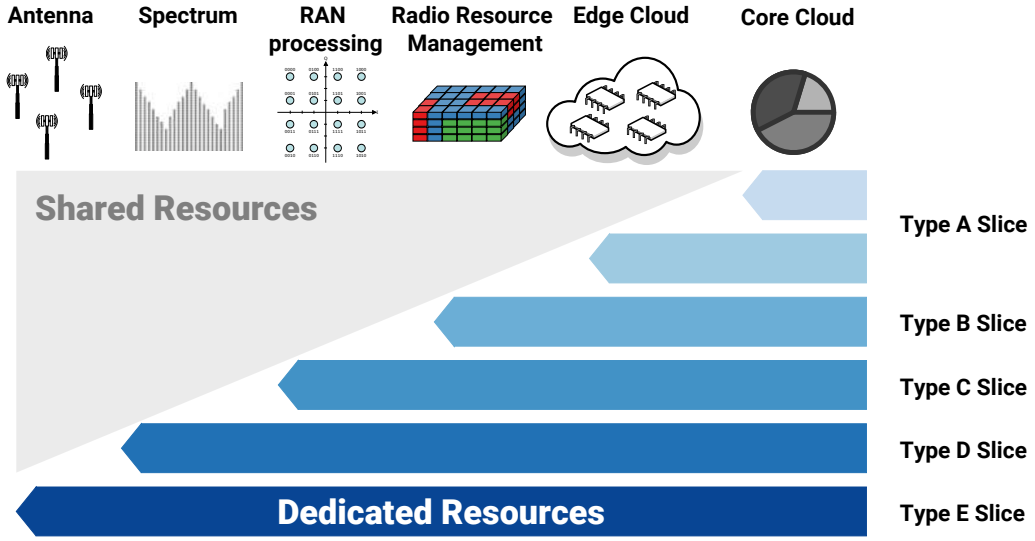


Figure 2.1. Network slicing strategies.

resource management efficiency, which increases by dynamically sharing the resources of the common infrastructure among the different services and slices; and, *(iii) system complexity*, resulting from deploying more dynamic resource allocation mechanisms that provide higher efficiency at the cost of employing elaborate operation and maintenance functions [40].

The above trade-off is fundamentally affected by the strategy adopted to implement network slicing, as illustrated in Figure 2.1. In its simplest realization, slices are limited to the core network (*type-A* slice in Figure 2.1): the allocation of resources to slices only involves cloud resources, and mostly becomes a VM or container resource assignment problem [32]. In this case, the level of service customization granted by slices is low, since it is restricted to core network functions; yet, high efficiency can be achieved at low complexity, as a large portion of the network remains shared among all services and tenants.

More dependable slicing would offer customized functions, possibly involving dedicated resources, also at the radio access, through, *e.g.*, Cloud/Centralized Radio Access Network (C-RAN) paradigms. Here, basic radio-access slices allow for tailored MAC-layer scheduling [41] across a large number of antennas (*type-B* slice). Moving down the protocol stack, advanced slices implement customized baseband processing (*i.e.*, encoding and decoding operations) in the Base Band Units (BBUs), possibly providing tenants with a guaranteed bandwidth at the air interface (*type-C* slice). These approaches provide the ability to customize scheduling strategies, but at the same time they reduce the possibility of radio resource sharing and/or increase the system complexity.

At fronthaul, resource isolation becomes a hardware problem [42]. A first case for slicing is one where tenants share antenna sites but are granted their own dedicated spectrum (*type-D* slice); we have virtually independent protocol stacks and full isolation, and sharing is limited to the physical hardware. Otherwise, tenants may require dedicated end-to-end resources down to the antennas (*type-E* slice); this results into slices that tell apart full, end-to-end virtual networks.

In general, slicing strategies at the higher network layers provide a lower level of customization yet they can more easily achieve efficient resource sharing without additional complexity. Indeed, when slicing occurs at high layers (*e.g.*, *type-A*),

the operator cannot offer full customization, but can easily employ highly dynamic allocation schemes for the lower layers; in contrast, achieving such an efficient resource allocation is much more challenging when considering network slicing schemes with stringent customization requirements (*i.e.*, strategies involving the lower layers down to *type-E* slicing). For instance, when all slices have a common MAC layer, an efficient sharing of radio resources is easy, yet MAC is not tailored to their different needs; conversely, if each slice implements a different, customized MAC protocol, it is more difficult to efficiently share radio resources.

However, the implications of network slicing in terms of efficiency of network resource utilization are still not well understood. Efficiency intuitively grows as one moves away from the radio access infrastructure (*type-E* slicing) towards the network core (*type-A* slicing); but we lack any more detailed characterization of the aforementioned trade-offs between customization, efficiency, and complexity. This is an important gap, since insights on the efficiency gains in network slicing are crucial to take informed decision on resource configuration strategies: if efficiency is preserved with solutions that assign resources to slices more or less statically, high customization levels can be achieved at a reduced complexity; however, if the price in efficiency is high, more elaborate (and expensive) solutions may be desirable.

To the best of our knowledge, this thesis presents the first work tackling the empirical assessment of network slicing in real-world networks. We believe that the insights it provides can be used as rule of thumb to evaluate the solution space for smart resource assignment algorithms and infrastructure dimensioning.

For instance, our results show that efficiency gains are very high in the edge, where employing technologies that allow for dynamic resource allocation provides a high reward; in contrast, gains are much reduced in the core, where complex, highly flexible reconfiguration schemes may not always pay off. Mobile network operators should thus be aware that isolating slices at the radio access may have a high cost in terms of efficiency, and that network slicing should be combined with solutions for dynamic orchestration of resources, at least at the network edge.

2.3 Related work for data analysis

As described in Chapter 1, related work that analyzes real-world mobile traffic data is scarce. In fact, the vast majority of analyses of cellular traffic builds on measurements and accounting data of voice calling and texting activities, such as Call Detail Records (CDRs); a thorough review is in [43]. Also, a number of works have investigated the properties of mobile data traffic from a high-level perspective, aggregating the load of all services [44–46]. The approaches above allow inferring important information on the communication patterns of users and on the overall data traffic they generate, but clearly do not explore subscribers' behavior on a per-service basis.

The literature becomes fairly thin when considering the usage of specific mobile services. Previous works have almost exclusively addressed the traffic dynamics of broad service categories (*e.g.*, *video* or *chat*) that encompass all mobile services of a kind. Service categories were proven to display interesting properties, including strong locality [17, 47, 48], high predictability [49], and adoption by well-outlined user groups [50, 51]. However, such broad categories hide the peculiarities of each service that, as we show in the following sections, are not negligible and deserve a dedicated investigation.

Considering individual mobile services (*e.g.*, *YouTube* or *WhatsApp*), they have been studied in isolation [52, 53], or within the scope of single categories such as video streaming [54] and mobile cloud [55]. As far as we know, the only work to consider a huge number of heterogeneous mobile services at once is that in [22]. There, the aim is comparing cellular and wireline traffic statistics for a relatively small (20-50,000) user population: both the purpose and the scale of the analysis are sensibly different from ours.

In contrast to the literature, in this work we analyze fine-grained service consumption from a European country. The dataset we employ in our study was collected in the core network of Orange, a major European mobile operator with a national market share of around 30%. We describe in each chapter the particularities of the studied data, since the relevant details vary as we focus on different aspects of the data according to the goal of each chapter.

2.4 Privacy, ethics and legal issues

We remark that our study abides by high ethical standards. All data collection, processing, and storage procedures at Orange were carried out in compliance with applicable regulations, including the European Commission (EC) General Data Protection Regulation (GDPR) [56]. These activities were supervised by the Orange Data Privacy Officer (DPO) as well as by the Commission Nationale de l'Informatique et des Libertés (CNIL), the French national body ensuring privacy in personal data use. Researchers outside the Orange premises only had access to traffic volumes aggregated at the antenna level over hundreds of users, which do not qualify as personal data and do not entail privacy risks.

2.5 Network security

Experts claim that network security is not horizontal anymore, but rather orthogonal to the network topology. Classical networks only allow to apply network security at the Layer 3 boundary (*e.g.*, using a firewall or router), but recent techniques such as Virtual Local Area Network (VLAN) stitching have enabled the use of transparent firewalls. In addition, using OpenStack Neutron networking [57] makes possible to provision security zones and protect application tiers using security groups rather than dedicated Layer 2 networks .

We allege that the the isolation between slices is the critical security issue related to slices. Nevertheless, Next Generation Mobile Networks (NGMN) already mention a list of key security concerns beyond isolation in [58]. Additionally, the work in Technical Reports [59, 60] undertakes security areas in 5G networks, and a full foresight of this dimension is addressed in [61].

5G architectures will have to provide at least the same security features than the previous Long Term Evolution (LTE) technology (*e.g.*, privacy and integrity in telecommunications, robust NFs, controlled access to admissible users). Besides, the virtualization framework has some risks associated, as compromised VNFs can affect others through the hypervisor, not only due to software bugs, but also because of hardware design failures.

Since monitoring or reporting slice security is an open field, it is still a challenge to ensure trust and consistent security policies between providers and tenants.

Nonetheless, in addition to other underlying framework, a dedicated slice for "control and management" of slice inter-dependencies may be needed [62].

As a remark, we provide these references in case the reader is interested in the security aspects, since security aspects are beyond the scope of this thesis.

3 Spatio-temporal analysis

Contents

As a first step towards understanding and exploiting data collected by mobile networks, we are interested in comparing the traffic demand patterns of different mobile services, exploring eventual common behaviors at a national scale and in two residential areas. Previous research on this topic unveiled that mobile apps tend to yield fairly comparable geographical pattern of consumption at national or regional scales [63], as we will observe. However, the same does not hold for temporal dynamics: apps have very diverse time series and even apps belonging to the same class feature unique combinations of activity peaks. This makes any attempt at clustering mobile services along a time dimension, as we will see in Section 3.3.

The characterization of mobile service consumption carried out in this work is also relevant to disciplines beyond networking. In fact, it allows observing social phenomena at unprecedented scales, unveiling interplays between the digital and physical worlds that are relevant to, *e.g.*, urban development [10] or planning [5, 11].

Hence, in this chapter we analyze fine-grained service consumption from a nationwide dataset. Our analysis is divided in four sections. First, we describe the dataset in Section 3.1, then we perform an analysis of all the services behavior 3.2. In Section 3.3 we explore the service time dynamics, and finally we present the geographical insights in Section 3.4.

3.1 Measurements and dataset

The dataset we employ in our study was collected in the core network of Orange, a major European mobile operator. It describes the mobile traffic generated by the whole Orange subscriber base in France, *i.e.*, a user population of approximately 30

million individuals distributed over more than 550,000 km^2. The data cover one week, starting on September 24, 2016. The time frame allows capturing the vast majority of mobile traffic dynamics, which are known to occur over weekly timescales [49,50], while avoiding that the dataset size becomes unmanageable. To maximise the generality of our results, the measurement week was carefully selected so as to avoid major nationwide events like holidays or strikes.

A simplified representation of the Orange 3G/4G mobile network architecture is portrayed in Figure 3.1. The Figure is limited to the 3G Universal Mobile Telecommunications System Terrestrial Radio Access Network (UTRAN) and packet switched core, and to the 4G Evolved Universal Mobile Telecommunications System Terrestrial Radio Access Network (EUTRAN) and evolved packet core, as our focus is on data traffic. The data were recorded by passive probes at the Gn and S5/S8 interfaces of the Gateway GPRS Support Node (GGSN) and of the Packet Data Network (PDN) Gateway (P-GW). The 3G and 4G gateways are conveniently co-located, which eases the probe deployment, management and synchronization. The probes inspect IP traffic on the GPRS Tunneling Protocol User (GTP-U), and extract information on the transport- and application-layer protocols of each user session. The specific mobile service associated to each IP session is detected by the mobile network operator via Deep Packet Inspection (DPI) and multiple fingerprinting techniques, each tailored to a specific traffic type. These operations can classify 88% of the mobile traffic; however, their implementation is proprietary to the mobile network operator, which prevents us from providing further details[1].

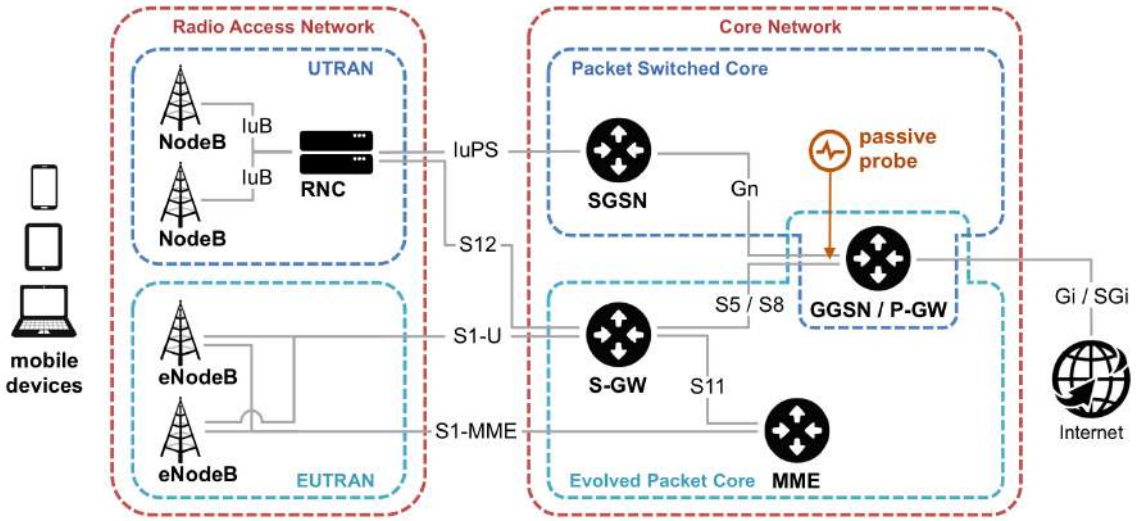


Figure 3.1. Simplified 3G/4G mobile network.

Geo-referencing of the IP sessions, and of the corresponding mobile service usages, is performed by examining the User Location Information (ULI) contained in the 3G Packet Data Protocol (PDP) Contexts and 4G Evolved Packet System (EPS) Bearers. These data structures are transferred over the GPRS Tunneling Protocol Control (GTP-C), which also transits through Gn and S5/S8 interfaces, making their inspection straightforward. The localization granted by this approach is fairly coarse, since the ULI is updated upon possibly infrequent events, *i.e.*, the establishment of

[1]Due to similar confidentiality restrictions, we do not disclose the absolute values of traffic volumes in the document, and limit our analysis to percentages of the total demand.

a new IP session, and handovers across access technologies (2.5G, 3G, 4G, Wireless Fidelity (Wi-Fi)) or Routing/Tracking Areas (RA/TA).

This limits the spatial accuracy of the dataset. As prior analyses showed that the median error of ULI is around 3 km [17], in our study we consider an appropriate tessellation of space at the level of *communes*. These are over 36,000 administrative regions whose union covers the whole France, and whose average surface is around 16 km^2. We thus associate each base station to the commune where it is deployed, and aggregate at the commune level all traffic mapped by ULI to base stations in that commune.

An issue with this process was that a number of rural communes have in fact no associated base stations, since coverage is provided by base stations deployed in neighboring communes. To avoid biases due to gaps in the spatial distribution, we evened out traffic in rural areas via Stewart potential model [64]. The model, commonly used in geography to estimate the reciprocal influence of spatial areas, was run on the set of communes with none to three base stations. The resulting redistribution effectively poises demands across rural areas, while conserving the global traffic volume of each service in such regions.

3.2 Mobile services overview

The dataset contains information about over 500 mobile services that generate some traffic during the measurement period. Figure 3.2 shows their ranking on the normalized traffic volume, in downlink and uplink. In both cases, rankings for the top half of services fit Zipf's distributions with similar parameters, at 1.69 and 1.55 respectively, before a cut-off intervenes that separates the bottom half of services.

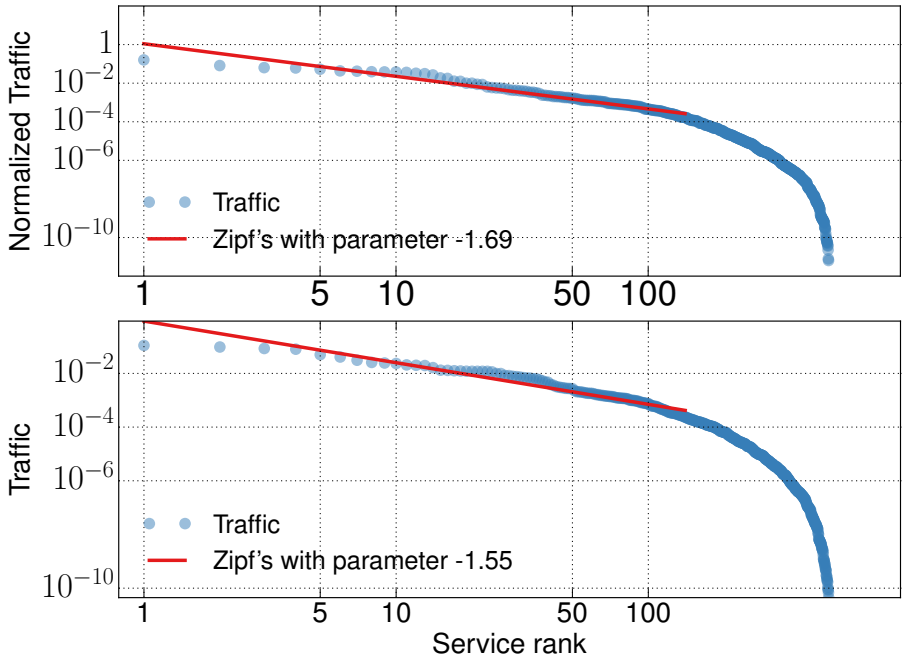


Figure 3.2. Ranking of mobile services on downlink (top) and uplink (bottom) traffic volume.

When comparing this plot to the equivalent one in [49], referring to 3G traffic measured in a US state in 2010, we remark that: *(i)* as in [49], the per-service traffic volumes span around 10 orders of magnitude, denoting a strong imbalance among

service loads; *(ii)* however, the distribution of traffic volume differs significantly from that in [49]: as mentioned before, only the top half of the services follows a Zipf distribution, and furthermore this Zipf distribution exhibits much lower parameters than the 4.74 value recorded in [49]. This latter observation highlights how per-service mobile traffic has evolved over the past six years: top-ranked applications now share more evenly the global demand, but a large number of very low-traffic services has also emerged.

In this study, we focus on the head of the distribution, which is composed of 20 representative services summarized in Figure 3.3. This subset of mobile services covers a large fraction (over 60%) of the overall network traffic and spans across a variety of service categories with diverse requirements in terms of network performance, such as *video streaming*, *gaming*, and *social networks*. The plots list the services (names on bars), categorized (colors of bars, as per the legend) and ranked on the relative traffic volume they generate.

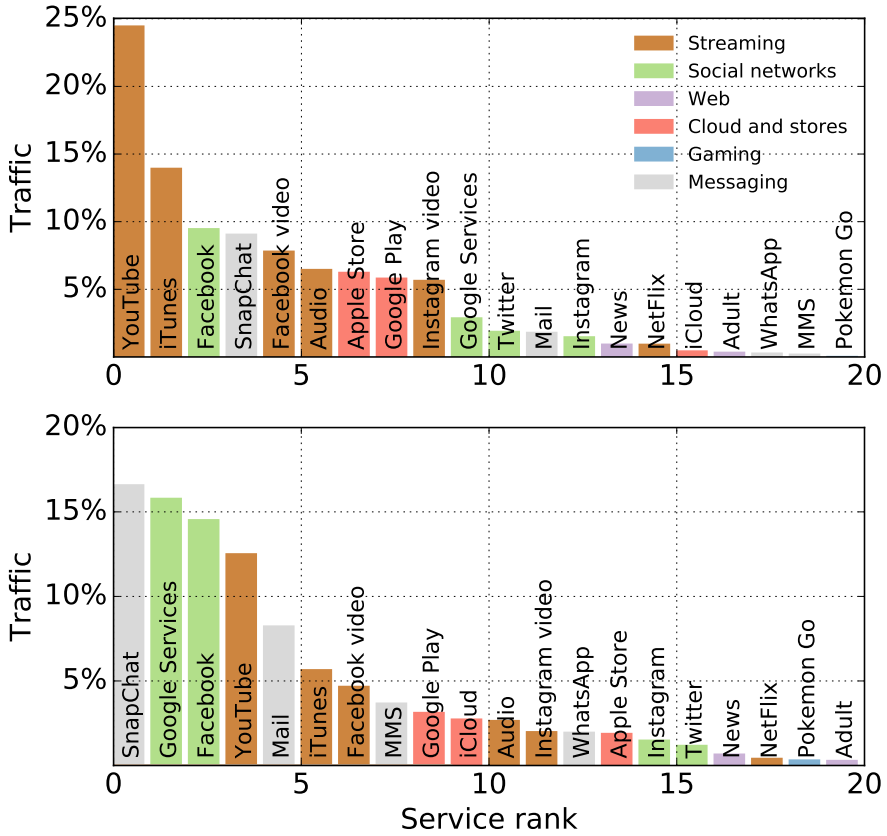


Figure 3.3. Selected mobile services, ranked on downlink (top) and uplink (bottom) traffic volume.

In downlink, video streaming services tend to dominate mobile downloads, with an aggregate figure at over 46% of the total traffic. This is a non-negligible increase over the 36% performance recorded six years ago in downstream cellular traffic [54]. It is also quite a different value from the 60% reported by Cisco[2] in their yearly forecast [65], for both cellular and Wi-Fi traffic: this lets us speculate that subscribers

[2]Cisco statistics refer to the aggregate downlink and uplink traffic. However, the latter accounts for less than one twentieth of the total network load in our case, hence it does not affect our conclusions.

may drastically reduce access to video streaming services when Wi-Fi is not available. YouTube emerges as the dominant provider, followed at a distance by iTunes.

Things change significantly in the uplink direction. Here, social networks and messaging services occupy the top three positions. This is not surprising, since services such as SnapChat and Facebook are oriented at content-sharing within limited circles (*e.g.*, of friends or contacts). Such a small potential audience of the content reduces the number of visualizations for the published material, ultimately resulting in high upstream-to-downstream traffic ratios.

3.3 Nationwide time dynamics

Our first goal is investigating the temporal dynamics of different mobile services. We focus on traffic at the national scale, aggregating the weekly demand for each service in space, over all communes.

Examples of the resulting time series are shown in Figure 3.4, for four sample mobile services in downlink. In all cases, classic patterns can be observed, *i.e.*, higher diurnal activity versus much reduced overnight traffic, and a distinctive dichotomy between weekends and working days. In addition, the time series of each service is characterized by a variety of fluctuations. For instance, the first plot in Figure 3.4, which refers to Facebook, displays a major traffic peak at midday of working days, plus several other minor peaks. However, other services in Figure 3.4 show other traffic peak arrangements.

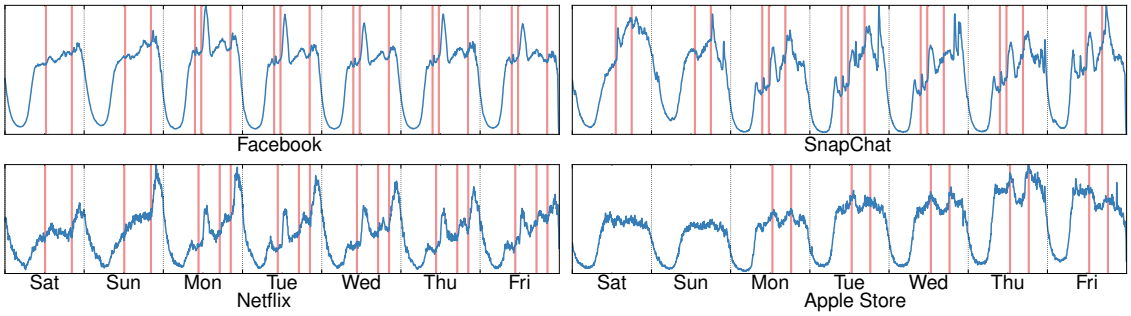


Figure 3.4. Sample time series of mobile services: vertical lines highlight activity peaks detected in the time series by the smoothed z-score algorithm.

Motivated by this last observation, we study how similar are the temporal patterns of mobile service usages in the whole France. Our aim is understanding whether the diversity of behaviors holds in general or is specific to the services considered in Figure 3.3.

We first make an attempt at grouping our 20 selected services based on similarity of their time series. Even if the original number of services is limited, a simple visual inspection of their time series can lead to subjective observations, challenging the validity and reproducibility of conclusions. We thus favor a sounder approach, and relay on a suitable clustering algorithm to carry out the classification task. Our choice is *k-shape*, which is the current state-of-the-art unsupervised technique for time series clustering, as proven by extensive tests [66]. We then carry out an exhaustive search for mobile service clusters, by testing k-shape on all possible values of k, in combination with multiple indices of clustering quality.

The latter are the Davies-Bouldin, modified Davies-Bouldin (top, minimum is best) and Dunn, Silhouette (bottom, maximum is best) indices , which constitute

a representative selection of popular indices used in the literature to rank different cluster sets generated from the same original elements [67]. Unfortunately, the outcome is inconclusive. As shown in Figure 3.5, none of the indices pinpoints a value of k as a clear winner. Instead, all indices indicate steadily decreasing clustering quality as k grows. Also, a thorough manual examination of the internal structure of the clusters generated by k-shape for different k does not reveal any consistent grouping of mobile services. We interpret these results as an indication that the temporal dynamics of our considered mobile services may be very distinctive, which makes them not easily equated.

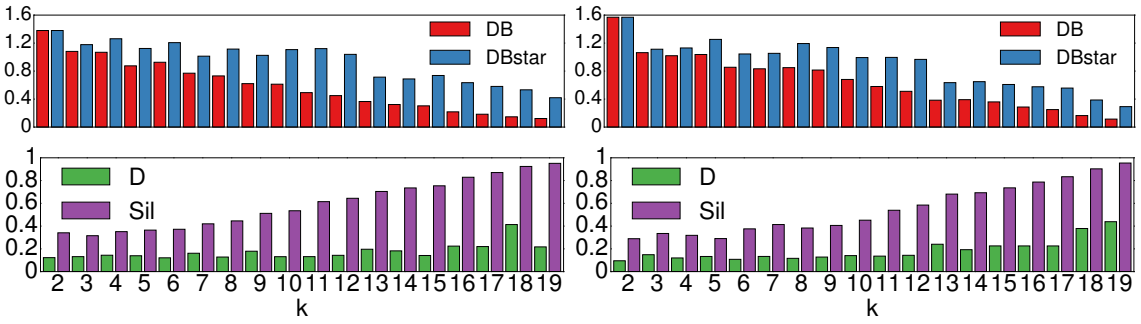


Figure 3.5. Clustering quality indices versus the cluster number, in downlink (left) and uplink (right).

The observations from Figure 3.4 suggest that the cause for the above behaviour may lie in the different patterns of activity peaks that characterize each service. In order to verify this possibility, we detect the activity peaks in the per-service time series using the *smoothed z-score algorithm*[3]. It compares the original signal versus a smoothed version of its z-score, and tags values higher than a threshold as peaks. It takes three parameters, controlling the threshold value (*threshold*) and the z-score smoothing, via the interval of past samples (*lag*) and their weight with respect to the current sample (*influence*). We set these parameters to 3 z-scores, 2 hours and 0.4, respectively, upon an extensive tuning process. An illustration of smoothed z-score peak detection is provided in Figure 3.6, for the Facebook case.

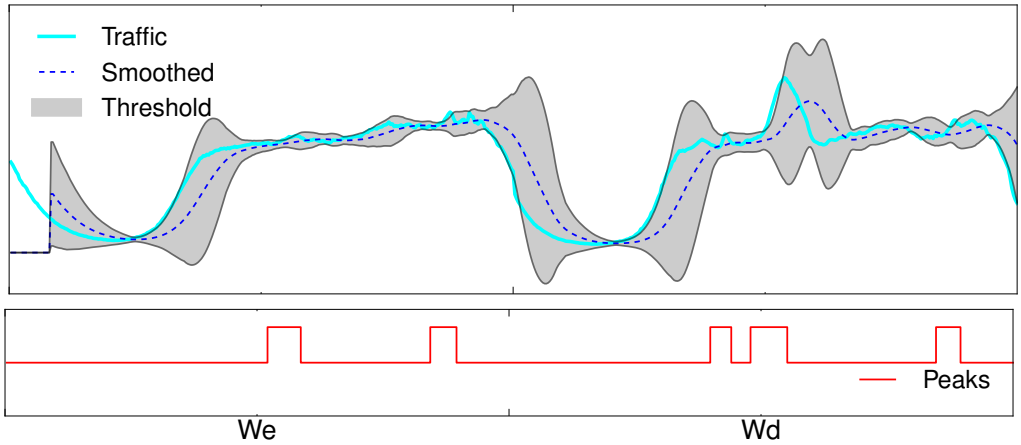


Figure 3.6. Example of smoothed z-score algorithm operation: Facebook.

[3]Implementation available at https://gist.github.com/ximeg/.

The top plot portrays the original signal, its smoothed version, and the range around it determined by the z-score threshold: when the original signal exceeds the range boundaries, a peak is detected in the bottom plot.

Examples of peaks inferred by the smoothed z-score are in the plots of Figure 3.4, where vertical red lines denote the rising front of peaks. Interestingly, by applying this methodology to all mobile services, we find that peaks only appear at seven specific moments during the week: at midday (around 1pm) and evenings (9pm) during weekends, and during the morning commuting time (8am), morning break (10am), midday (1pm), afternoon commuting time (6pm) and evenings (9pm) during working days.

This lets us summarize the peak patterns observed for all services as done in Figure 3.7. In this Figure, each sector refers to one mobile service, and each ring to a different topical time, as per the legend. We remark that: *(i)* individual services tend to have very diverse patterns even when looking only at when they show peaks of activity; *(ii)* this heterogeneity also separates services that belong to a same category, *e.g.*, video streaming behaves quite differently in YouTube, Facebook, Instagram, NetFlix and iTunes platforms.

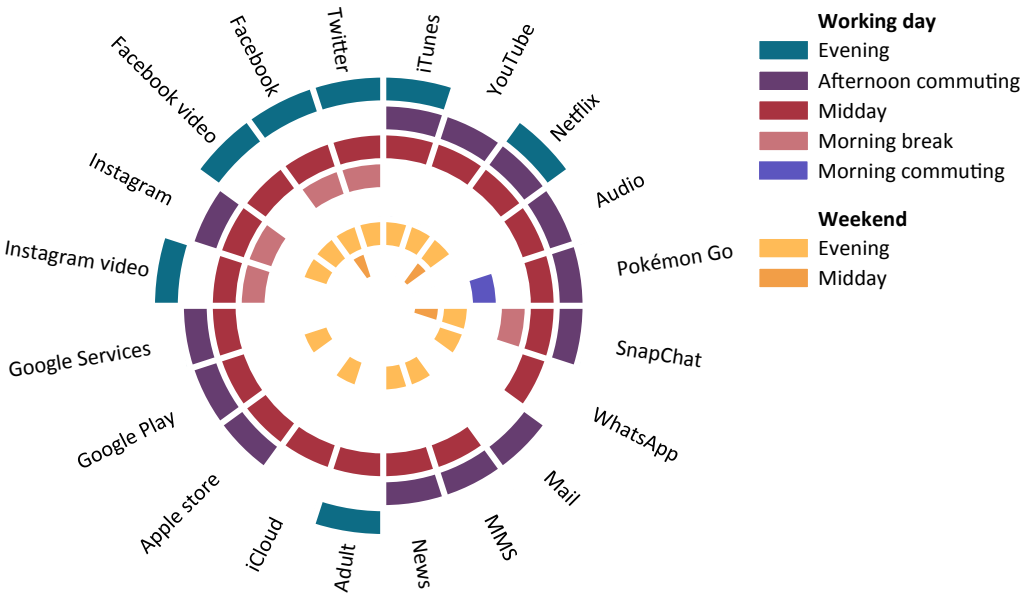


Figure 3.7. Activity peak times of mobile services.

Other specific behaviors have interesting implications. For instance, almost all services show increased usage on midday of working days. Similarly, large (different) sets of services have activity peaks during the afternoon commuting time and during weekend evenings. In all these cases, the increased usage affects services of various nature, indicating that subscribers with different interests all tend to consume mobile services at those times. On the contrary, we speculate that morning break activity peaks may highlight services that are popular among students, who access them during the pause between classes: coherently, these include SnapChat, Instagram, Facebook, and Twitter.

Figure 3.8 offers an in-depth view of the activity peaks: it displays, for each topical time, the actual activity peak intensity of every service. This is computed as the ratio between the maximum and minimum traffic volumes recorded during the peak

intervals as detected by the smoothed z-score algorithm. The key observation here is that services with demand peaks at a same time in fact undergo very diverse variations of activity.

Overall, the diversity of activity peaks, both in timing and intensity, corroborates the intuition that temporal fluctuations in the usage of individual mobile services are very heterogeneous. These results explain the poor outcome of a clustering based on time series, and ultimately demonstrate how each mobile service has unique dynamics, dictated by the classes of subscribers consuming it.

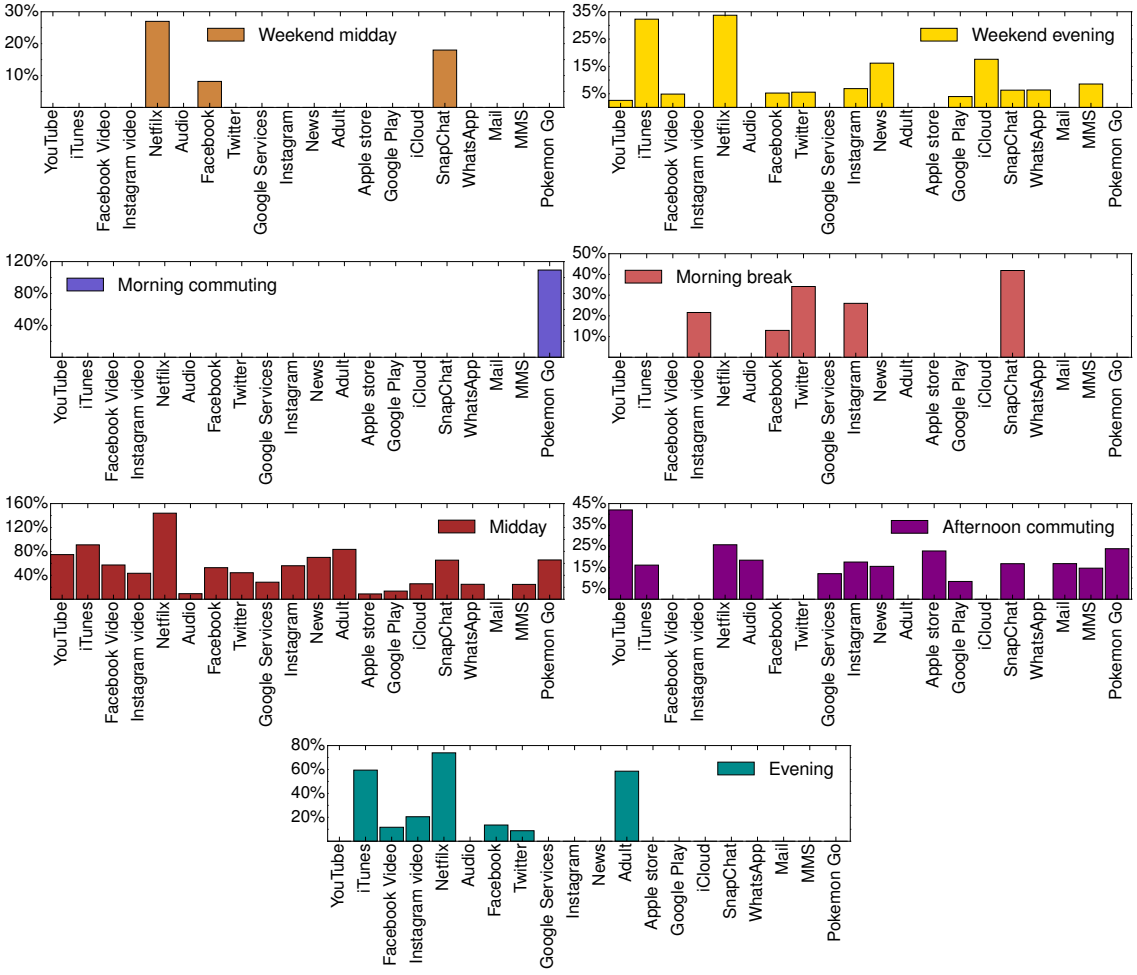


Figure 3.8. Peak-to-average ratios measured for each mobile service at different topical times.

3.4 Service usage geography

Mobile services show significant peaks of activity not only in time, but also in space. An example is provided in Figure 3.9 for the specific case of Twitter. The left plot shows the cumulative weekly downlink and uplink traffic recorded in the ranked communes. We observe that the top 1% and 10% of the communes generate over 50% and 90% of the Twitter traffic, respectively.

The strong imbalance in the demand recorded across communes is expected, considering the high heterogeneity of population density that characterizes France. What is less obvious is that such variability remains strong also when considering the average traffic generated by a subscriber, obtained as the ratio of the traffic volume to the average number of users in each commune. The right plot in Figure 3.9 shows the Cumulative Density Function (CDF) of the *per-subscriber* Twitter usage over all communes. The distribution is highly skewed: subscribers in half of the communes consume a negligible weekly Twitter load of a few 1 Kbytes, whereas users in other areas download tens of Mbytes of Twitter contents per week. Basically, individual mobile users who live in distinct areas of France tend to use Twitter services in very different quantity.

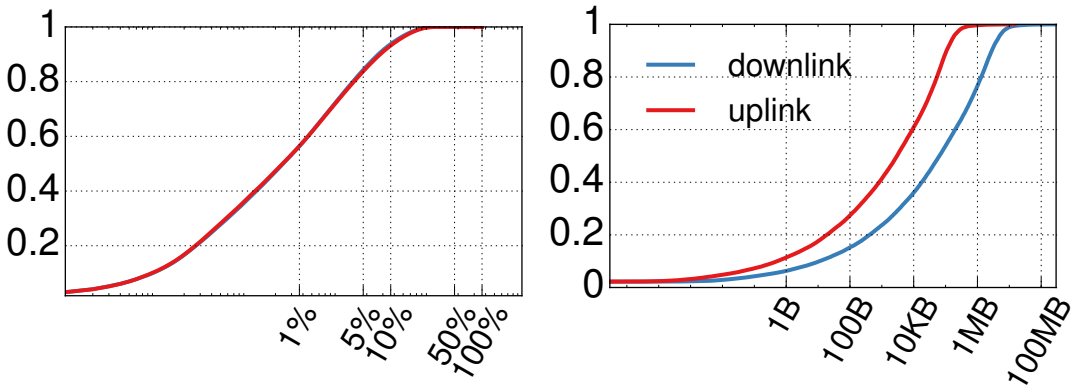


Figure 3.9. Cumulative traffic on ranked communes (left). CDF of per-subscriber traffic on all communes (right).

The left plot in Figure 3.12 reports a map of the weekly per-subscriber Twitter traffic in downlink, and helps to visualize this phenomenon. The map evidences how users living in large cities (*e.g.*, Paris, Lyon and Marseille) and traveling along major transportation arteries (*e.g.*, the high-speed "Train à Grande Vitesse" train (TGV) lines connecting the three cities above) tend to generate significant demands. Instead, subscribers in rural areas located far from major cities and transportation infrastructures are prone to make little use of mobile services.

The considerations above refer to Twitter, but they are valid for any mobile service. This is shown in Figures 3.10 and 3.11. The plot in Figure 3.10 is a CDF of the geographical correlation of mobile service usage: we represent each mobile service as a vector of the weekly per-subscriber traffic recorded in each commune, and compute the coefficient of determination between the vectors of each pair of services. The majority of values in the CDF are strongly positive, with an average of 0.60 and 0.53 for downlink and uplink respectively.

The plots in 3.11 detail the correlation among specific service pairs: low correlations are only experienced with NetFlix (almost completely absent in rural

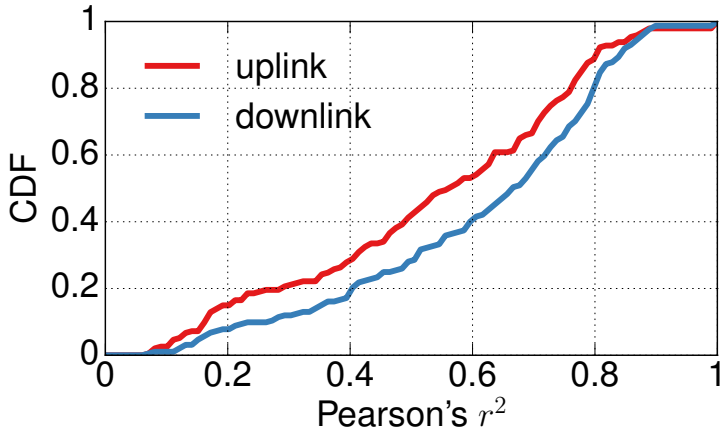


Figure 3.10. Pearsons' r^2 CDF computed between the per-user traffic maps of all service pairs.

areas) and iCloud (pushing uplink data from all iPhones, and thus more uniformly distributed over the country). These outlier cases apart, our results let us conclude that mobile services tend to be consumed similarly over the French territory (see Figure 3.12). Such mobile data consumption features geographical distributions of the per-user traffic that are highly skewed, as in Figure 3.9, with subscribers within cities and on inter-city routes that tend to be more active than those in other areas, as per the first plot (a) in Figure 3.12.

The second plot (b) in Figure 3.12 provides instead additional detail on the NetFlix outlier, showing the corresponding map of weekly per-subscriber traffic. Densely inhabited city centers and major transportation lines stand out much more clearly than in the typical case, in the left plot (a) of the same figure. This occurs also because NetFlix usage is dramatically low, or even absent, in large regions of rural France. We partly ascribe such a strong duality of adoption to the high-end nature of NetFlix as a mobile service, with users in cities more prone to embrace novel applications. In addition, the NetFlix case also gives us the opportunity to discuss how the mobile network technology is an important factor that can further explain the success of specific services. Streaming high-quality long-duration videos requires substantial capacity and quality of service, and indeed the 3G and 4G coverage in France, in the bottom-right plot of Figure 3.12, seems to drive NetFlix usage over the country. In the case of other services, such as Twitter, the spatial distribution is more uniform; when looking at the 3G and 4G coverage, this suggests that (pervasive) 3G already provides sufficient performance, and makes demands less dependent on the cellular technology.

In order to further explore the impact of urbanization on the way mobile services are consumed, we group communes in France into *urban*, *semi-urban* and *rural*, according to classifications of the French National Institute of Statistics[4]; in addition, we consider rural communes that are crossed by a high-speed train line into a separate *TGV* category. We then aggregate all traffic recorded in the communes of a same group, and investigate their relationships.

The top plot in Figure 3.13 summarizes, for each service, *how much* traffic is generated by the average individual subscriber in semi-urban, rural and TGV regions

[4] https://www.insee.fr/fr/information/2115011

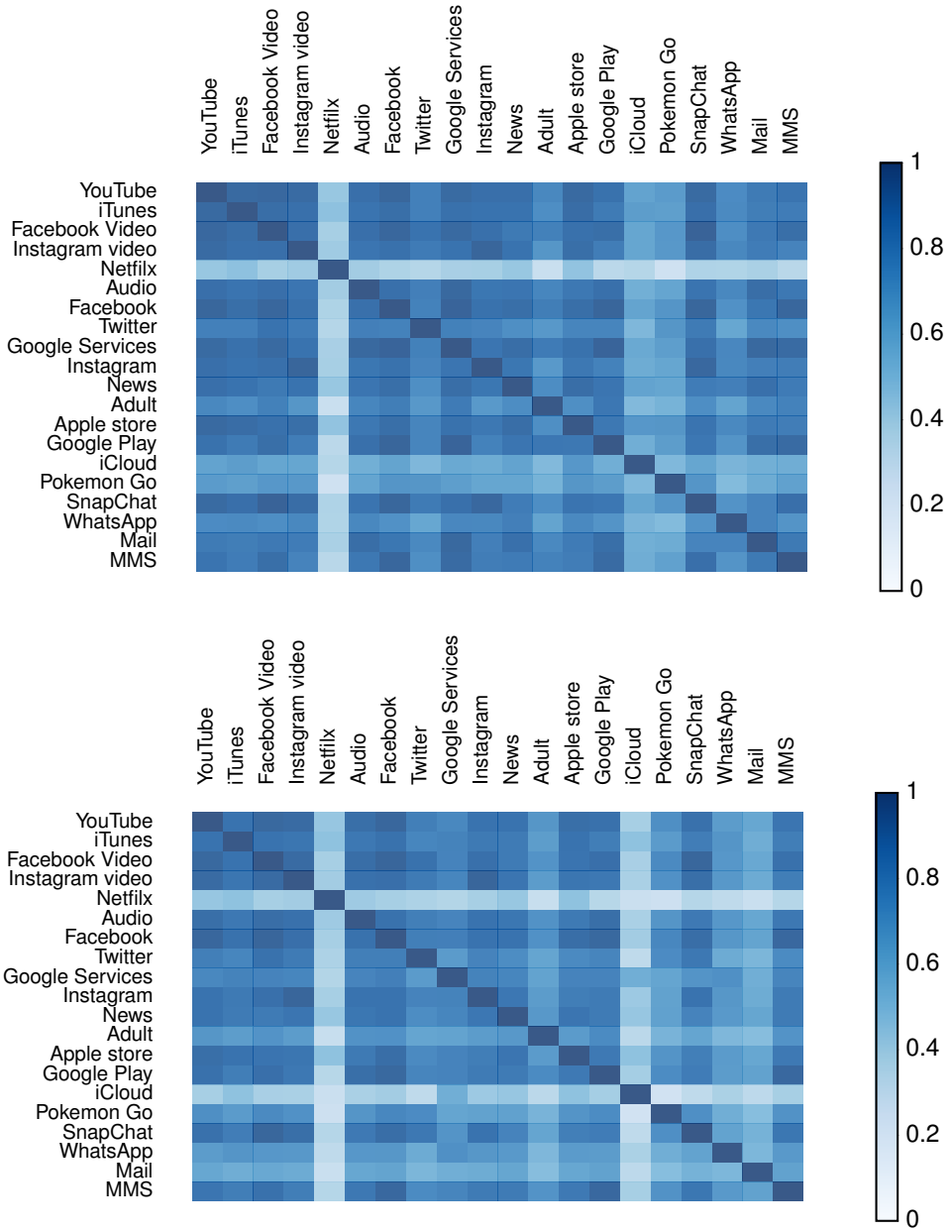
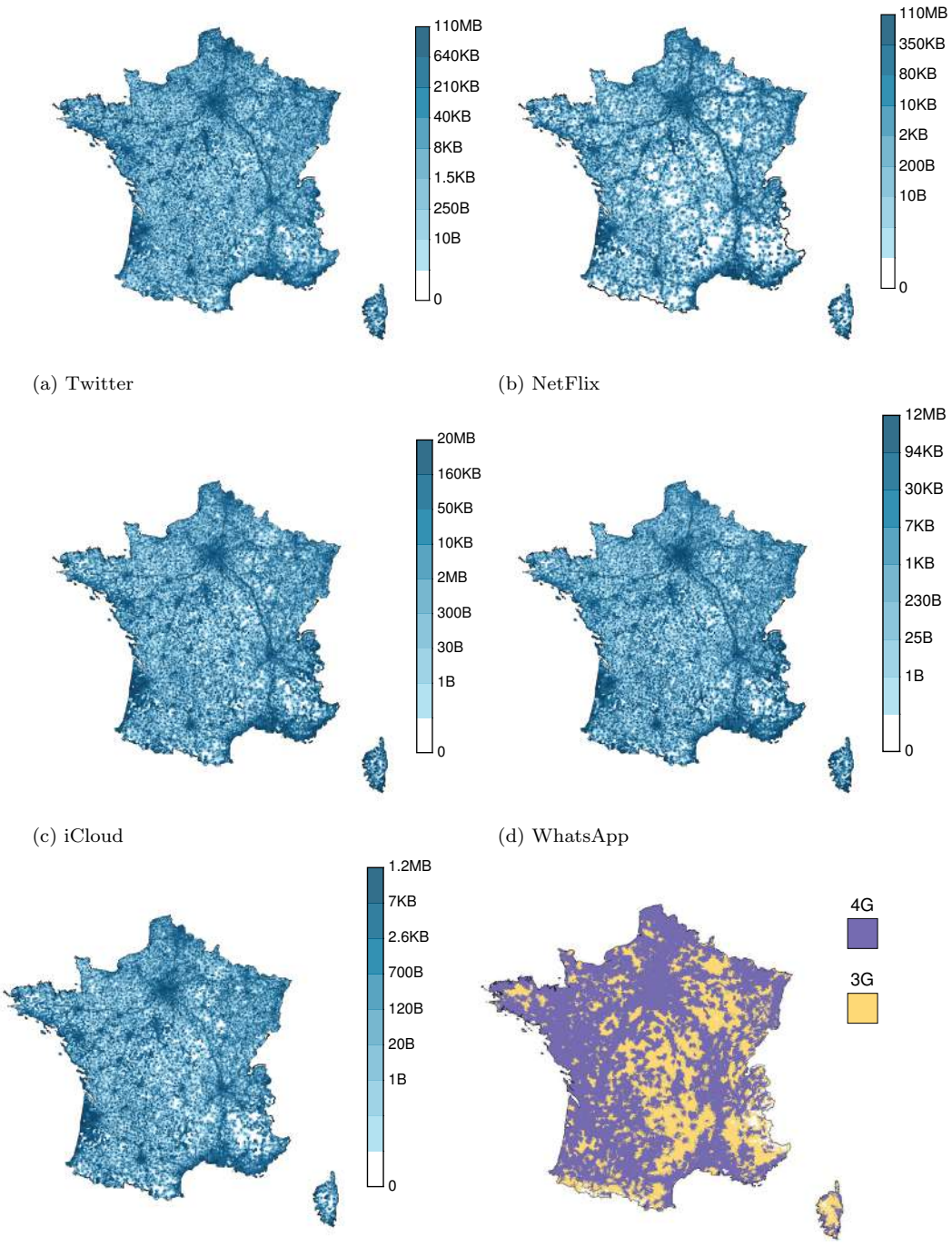


Figure 3.11. Pairwise coefficients of Pearsons' r^2 for downlink (top) and uplink (bottom).

with respect to a typical user located in a urban area of France. More precisely, each bar represents the slope of the linear least square regression of per-subscriber time series in urban and (from darker to lighter) semi-urban, rural and TGV regions. The plot highlights that: *(i)* semi-urban and urban areas present similar individual service usage levels, *i.e.*, the coefficient is close to 1; *(ii)* subscribers in rural areas consume around a half of the mobile service data than their counterparts in urban zones do; *(iii)* users on high-speed trains generate on average twice or more the volume of traffic of urban users.

Interestingly, all these results are fairly consistent across services. They unveil how users in cities, from medium-sized towns to large metropolis, show equivalent mobile services consumptions. On the other hand, urban mobile service consumptions are twice as large as those of people living in the countryside, which may be covered by



(a) Twitter

(b) NetFlix

(c) iCloud

(d) WhatsApp

(e) Pokemon Go

(f) Coverage of Orange 3G and 4G cellular technologies in France

Figure 3.12. Maps of the average per-subscriber activity for downlink traffic.

3G connectivity only. That trend is exacerbated for passengers on high speed trains, who are much more prone to use mobile services during their travels: rural communes belonging to the TGV category show completely opposite relative consumption trends when compared to non-TGV ones.

The bottom plot in Figure 3.13 assesses instead if the urbanization level plays a role in *when* the typical subscriber access mobile services. The bars represent the mean coefficient of determination between the time series of a same service recorded in one type of region and those of the other types. Across the vast majority of services,

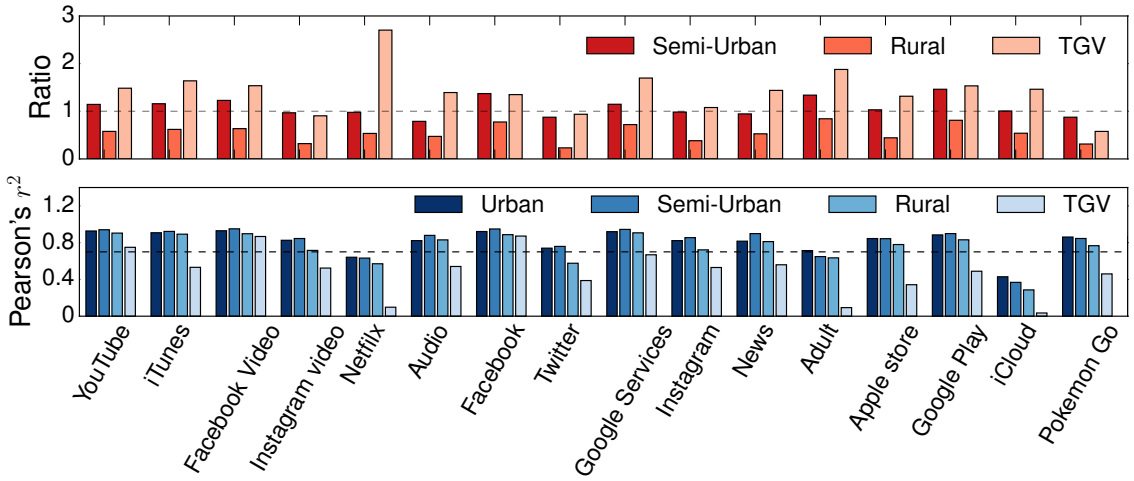


Figure 3.13. Per-user traffic volume ratios among urbanization levels (top). Correlation of per-user traffic time series among urbanization levels (bottom).

correlations are high for combinations involving urban, semi-urban and rural areas; thus, we conclude that the level of urbanization has very little impact on the temporal dynamics of service usage. Instead, subscribers on TGV have quite different temporal patterns than users in the rest of France. We argue that the cause is a combination of train schedules (constraining usages in time), and the intrinsic nature of some services (*e.g.*, TGV seats are probably not the best environment for adult websites browsing).

4
Spectral analysis

Contents

Our previous analysis about the time and spatial components of the time-series was useful to understand trends associated to mobile services, but a more automated way of classifying them may be derived based on a frequency analysis. In fact, we show that the spectral components reveal significant common properties in the time series of data traffic generated by many diverse mobile services. Such shared traits emerge in the form of periodicity in the demand fluctuations: we find that most apps have activity peaks occurring at similar frequencies. Therefore, our results indicate the existence of temporal regularities in the consumption of mobile services, which are most likely driven by the frequency of the same underlying human routines. As such, they are a first step towards a comprehensive classification of mobile service based on how they are used in time.

The chapter is organized as follows. We present the new data in Section 4.1, and detail preliminary denoising steps in Section 4.2. The spectral analysis of time series inferred from such data is in Section 4.3. Section 4.4 presents the results of a dedicated clustering of the harmonics returned by the spectral analysis.

4.1 Dataset

Our analysis relies on measurement data collected in the mobile network of Orange again, but in this occasion the data was recorded during three consecutive months in Fall 2016, following the same methodology as in Section 3.1.

The dataset employed in our study describes mobile service traffic generated in the metropolitan area of Paris, France. This is one of the largest conurbations in Europe, covering an area of over 100 km^2 and more than 2 million inhabitants. Orange had a 2016 market penetration of around 34% in the region, hence our data are representative of a large fraction of the local population. As aforementioned in the previous chapter, we remark that the localization accuracy of ULI is known to be limited due to irregular updates to the field; however, the typical precision, in the order of km [17], is sufficient to correctly locate the traffic produced within a large city such as Paris. The data have a temporal granularity of 5 minutes.

We focus our study on 37 mobile services. The rationale for this choice is that traffic volumes generated by mobile apps follow a well-studied power law [1, 49], hence only a very limited set of services yield considerable demands that are worth investigating. Our choice of mobile services includes heterogeneous applications that rank among the top 50 in terms of traffic load, and encompass video and audio streaming (*e.g.*, YouTube and iTunes), social media (*e.g.*, Facebook and Instagram), messaging (*e.g.*, Snapchat and WhatsApp), stores (*e.g.*, Apple Store and Google Play), gaming (*e.g.*, Pokemon Go and Supercell), as well as traffic generated by generic digital activities (*e.g.*, web browsing and electronic mail). Further details are in Section 6.5.1.

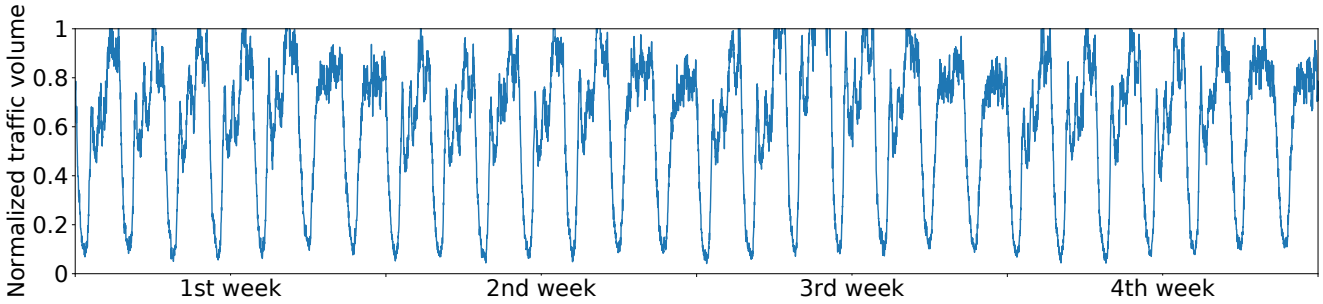
4.2 Preprocessing

Our analysis focuses on understanding *typical* patterns of mobile service demands. The raw time series of app traffic recorded over three months definitely capture such patterns, yet they also feature fast-varying noise (due to the inherent randomness of user access), and long-timescale trends (due to, *e.g.*, holiday periods or diverse daylight intervals). In order to filter out such phenomena, and work with cleaner time series, we preprocess the data, by computing a *median week* traffic demand for each mobile service [11].

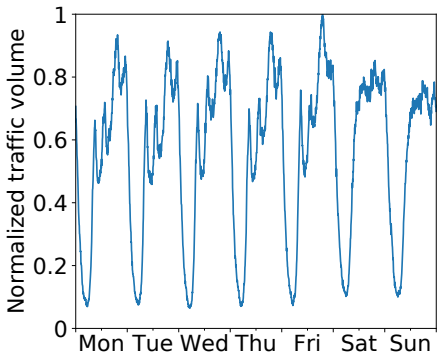
Let the demand for a given mobile service s be described by a three-month time series $d^s(t)$, where $t > 0$ has a resolution matching some time unit, *e.g.*, minutes. The median week is an ordered vector $\boldsymbol{w}^s = (w^s_n)$, $n \in [0, N-1 = \lfloor W/T \rfloor) \subset \mathbb{Z}$, where W is the duration of one week in the considered time unit, *e.g.*, 10,080 in minutes, and T is the duration of one time slot in the same unit. In our analysis we consider five-minute time slots, hence $T = 5$. Each element of $\boldsymbol{w}^s$ is expressed as:

$$w^s_n = \mu_{1/2} \left\{ \sum_{t=\tau}^{\tau+T-1} d^s(t) \;\middle|\; \frac{\tau \bmod W}{T} = n \right\}, \tag{1}$$

where the operator $\mu_{1/2}$ denotes the median of the value set in the argument. Equation (1) divides each week into slots spanning T, and computes the overall traffic observed for service s within each such slot; then, slot n of the median week $\boldsymbol{w}^s$ is assigned a single value w^s_n equal to the median of (sum) values associated with the n-th slot of each week.



(a) Four-week traffic time series



(b) Median week time series

Figure 4.1. Traffic time series generated by YouTube during four consecutive weeks, and corresponding median week.

The preprocessing above has the effect of eliminating noise and seasonal effects and returning a more accurate representation of the ordinary demand generated by each mobile service. Note that we use the median instead of other statistical measures since it is more robust to strong outliers that risk for instance to bias averages. Figure 4.1 shows an example of a three-month time series (a) in the target metropolitan area compressed into a median week representation (b).

4.3 Frequency analysis

We employ Fourier decomposition on the time series presented before. We remark that spectral methods have already been applied to mobile data traffic time series [5]; however, they were only used for data denoising and not for interpretation, and considering aggregate traffic instead of service-level demands. Our approach allows deriving the frequency spectra of mobile service demands, as outlined in Section 4.3.1, and discuss how they decompose into harmonics of different importance, as set forth in Section 4.3.2.

4.3.1 Service demand spectra

Fourier decomposition allows approximating complex time series as sums of simple trigonometric functions. Specifically, since our median week representations describe a discrete-time process over a finite interval, we apply a Discrete Fourier Transform (DFT). Given the signal $\boldsymbol{w}^s$ for service s, its DFT is a complex-valued function of

discrete frequency, $\boldsymbol{X}^s = (X_k^s)$, $k \in [0, N) \subset \mathbb{Z}$. The granularity of the DFT in the frequency space is the reciprocal of the duration of the time signal, hence $1/(NT)$ in our case. Then, the k-th component of the DFT describes the value of the function at frequency $k/(NT)$, as:

$$X_k^s = \sum_{n=0}^{N-1} w_n^s \exp\left(-i\frac{2\pi}{NT}kn\right). \tag{2}$$

The DFT is a linear and invertible operation, and the inverse DFT allows reconstructing the time series from its DFT. Formally, the value of the median week demand for service s at each time slot n can be derived from $\boldsymbol{X}^s$ as:

$$w_n^s = \frac{1}{N}\sum_{k=0}^{N-1} X_k^s \exp\left(i\frac{2\pi}{NT}kn\right). \tag{3}$$

Intuitively, upon inversion, each DFT component translates to a sinusoidal function of time with frequency k/NT, and amplitude and phase described by the phasor X_k^s. The time series of s is expressed in Equation (3) as the sum of such sinusoidal functions.

We compute the DFT of all mobile service demands, so as to reveal their underlying frequency components. The DFT returns N components for any given service s, where N is typically large; however, these components have very different X_k^s values. As mentioned above, these values embed the amplitude of the inverted sinusoidal function for component k, and only components with sufficiently high amplitude contribute in a significant manner to the original signal. In order to understand the importance of each component for mobile service demand, we resort to the DFT power spectrum, which is computed as $|\boldsymbol{X}^s|^2$ and describes the distribution of power across frequency components.

Figure 4.2 shows a representative subset of the power spectra for a selection of services. The spectra are centered at zero frequency, and the central value $|X_0^s|^2$ is the power[1] associated with the constant mean of the time series $\boldsymbol{w}^s$. As one moves away from the central value, the spectra portray the power of increasingly higher frequencies, up to $(N-1)/NT \sim 1/T$ for large N. All plots in Figure 4.2 have very similar shapes that highlight how the power spectra of median week traffic demands are dominated by low-frequency components for all services: as highlighted by the logarithmic ordinate, central frequencies have much higher $|X_k^s|^2$ values. This implies that the traffic demand for any service is strongly characterized by regular patterns with periods in the order of hours and days.

Faster dynamics with periods in the order of tens of minutes or less have a much reduced power. However, we also remark that the spectra flatten around the central peak, indicating that the original signals are also fairly noisy, and all high-frequency components are needed to perfectly reconstruct the time series.

[1]Note that power values in all figures are normalized with respect to the total signal power, so as to make them more easily understood.

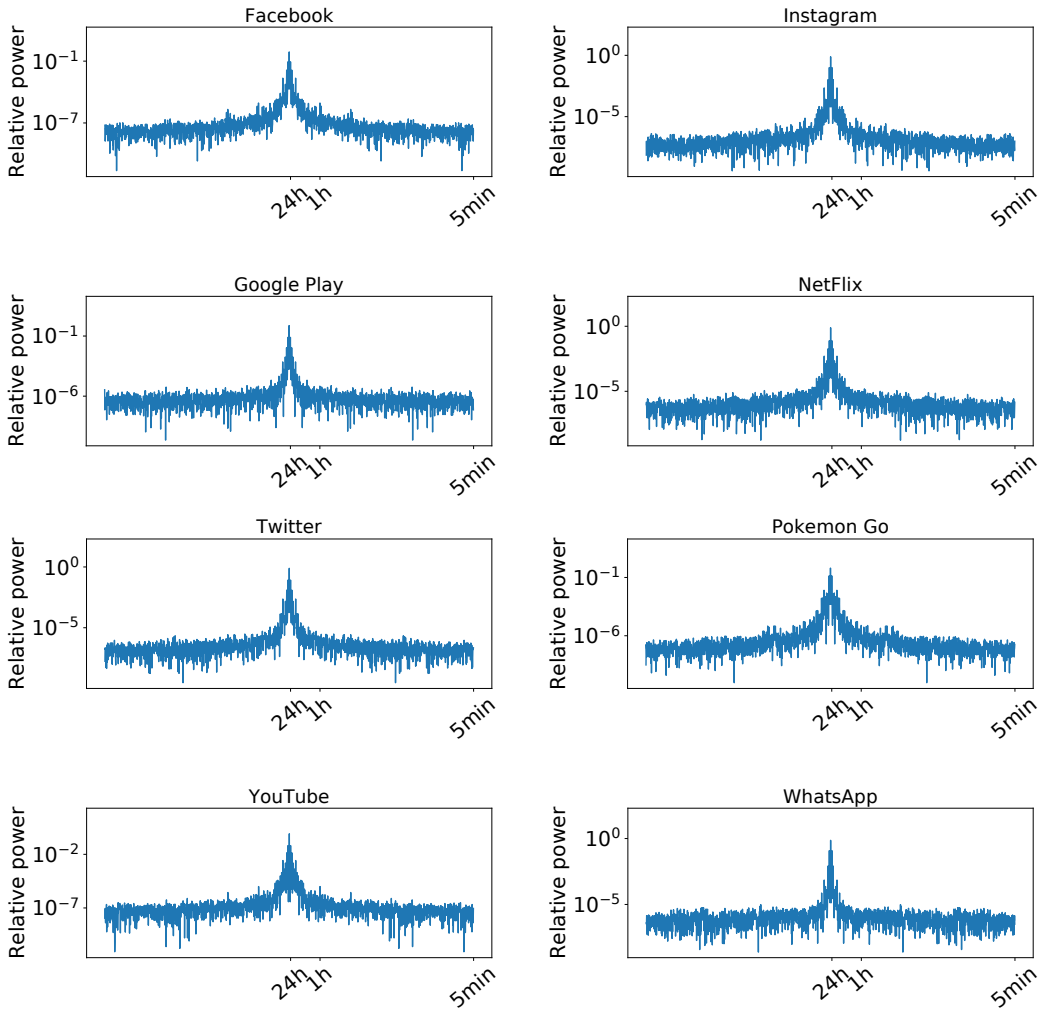


Figure 4.2. Power spectra of the DFT of selected mobile services.

4.3.2 Component filtering

In order to provide better insights on the most important components, Figure 4.3 offers a detailed view of the central frequencies of the spectra in Figure 4.2, for $k \in [0, 50]$. This is equivalent to looking at periodicities NT/k of roughly 4 hours or longer, as marked on the abscissa. These plots make it clear that diversity exists even among central components: for each service, specific frequencies have peaking $|X_k^s|^2$ values, *i.e.*, are especially critical to the temporal dynamics of the demand.

Given that components have heterogeneous power, and the vast majority only marginally contributes to the original signal, it makes sense to limit our analysis to a subset of relevant components for each service. To this end, we retain the minimum number of components whose summed $|X_k^s|^2$ values preserve at least 99% of the total signal power, excluding all components whose contribution is below 0.1%. This is equivalent to ranking components for each service based on their associated power, in descending order, and then computing the cumulative sum of such power following the ranking.

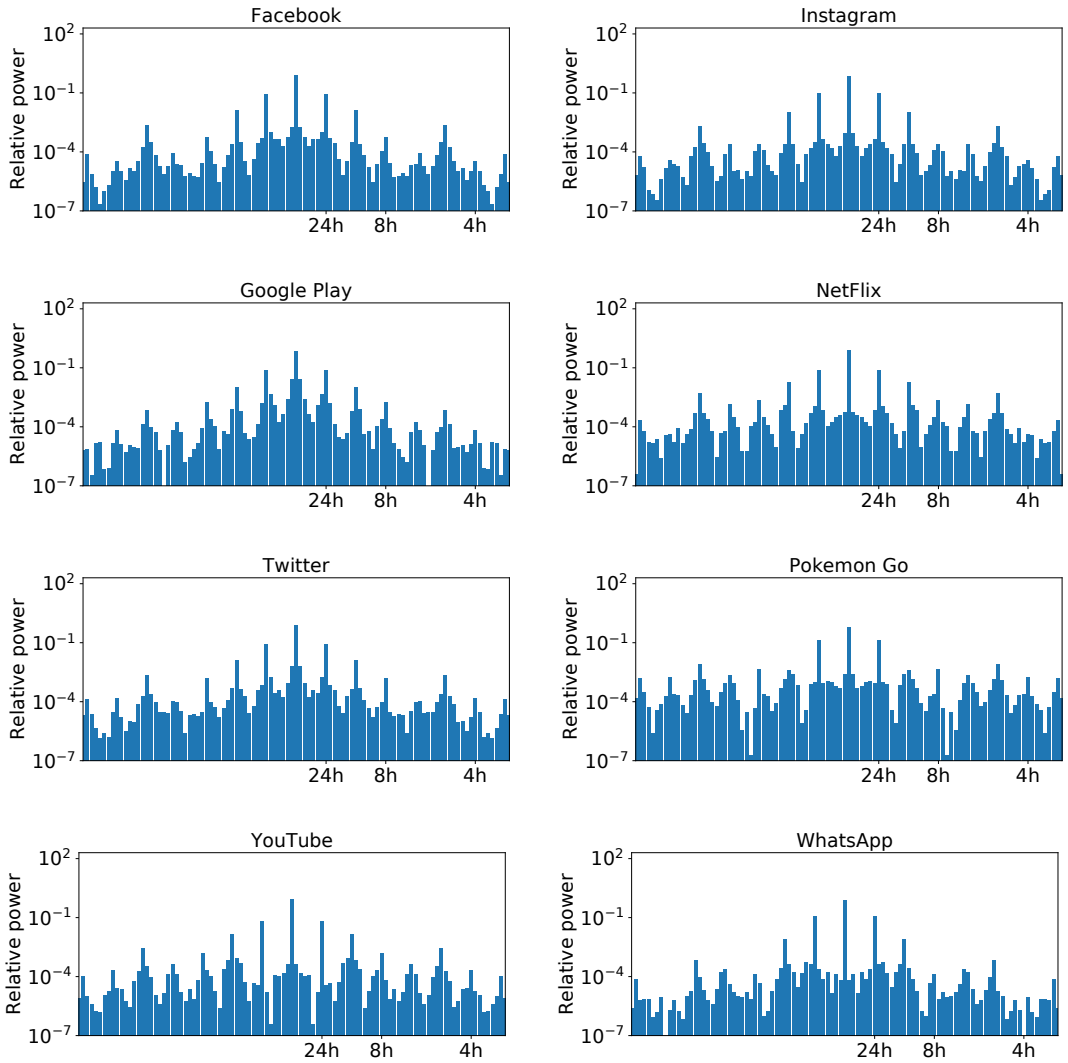


Figure 4.3. Low-frequency components ($k \in [0, 50]$) of the power spectra of the DFT of the selected mobile services in Figure 4.2.

Figure 4.4 illustrates the result for the services in Figure 4.2. The component ranking first is invariably that associated with the mean of the time series and accounts for the vast majority of the signal power. More importantly, the number of additional components needed to attain the 99% threshold is small in the vast majority of cases.

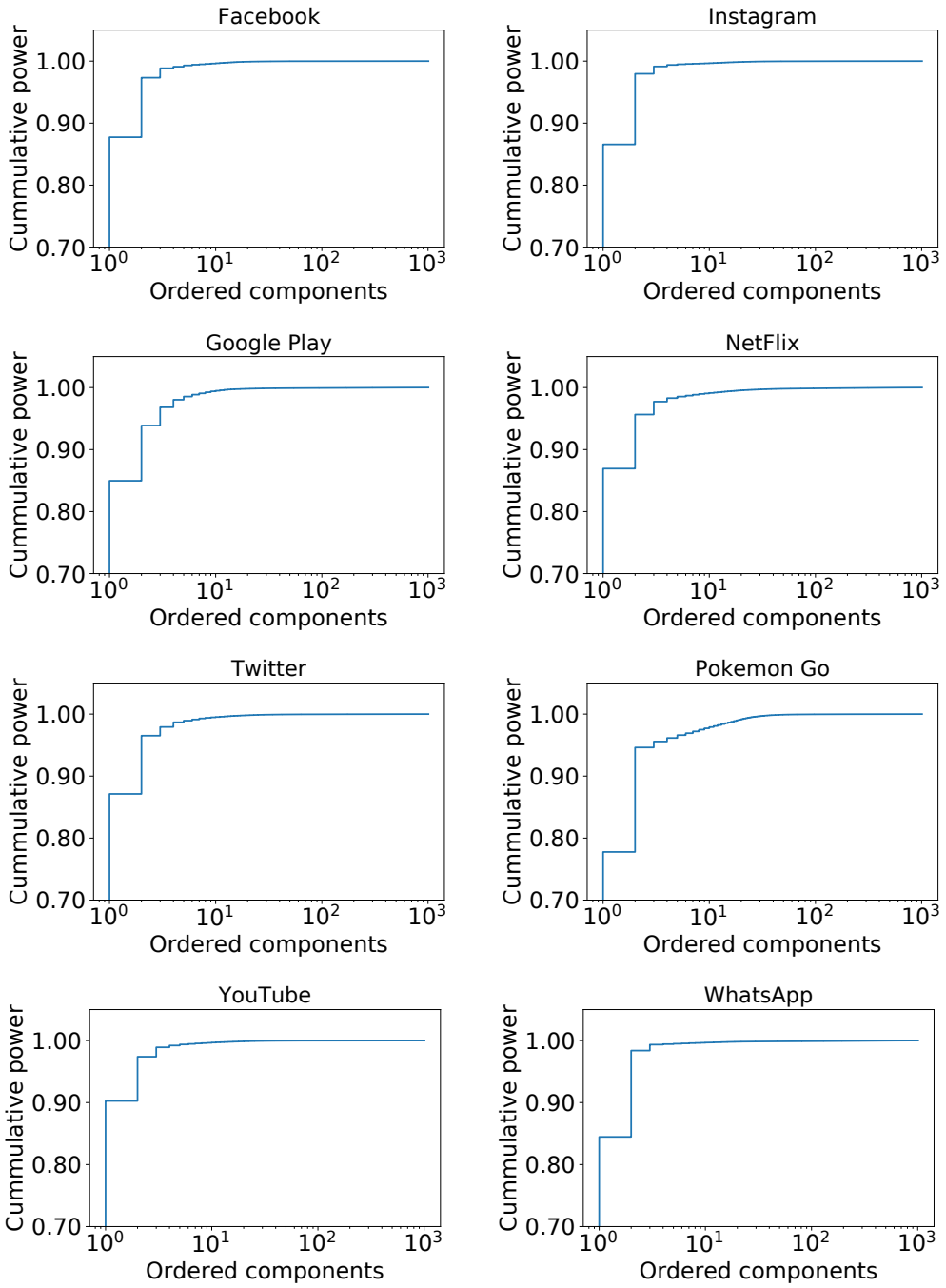


Figure 4.4. The cumulative sum of the power associated to ranked components, for the selected mobile services in Figure 4.2.

Table 4.1 summarizes the number of components needed to reach the target cumulative power, for all mobile services. Less than 10 components are sufficient to attain the 99% threshold in almost all cases, and the traffic demands of many popular applications such as YouTube, Instagram or Facebook are mostly influenced by recurring patterns at five or six different frequencies. In light of these results, in the remainder of our analysis, we focus on the high-power frequencies listed in the Table only and discard all other noisy components. Also, we do not consider the zero-frequency components in subsequent discussions, since the mean is a service-specific constant value that only captures the volume of traffic associated to each service, and is irrelevant to the temporal dynamics we are interested in. Ultimately, we retain 326 components across all service traffic demands for further analysis.

Services	# Components	Retained Power
WhatsApp (m2)	3	99.34 %
MMS (m4)	4	99.51 %
iCloud (c3)	5	98.87 %
YouTube (s1)	5	99.37 %
Generic messaging (m5)	5	99.42 %
Instagram (n4)	5	99.49 %
Instagram video (s4)	5	99.57 %
News (w3)	6	98.79 %
Generic video (s7)	6	99.32 %
Facebook (n1)	6	99.42 %
Google Services (n3)	6	99.42 %
Ads (w1)	6	99.43 %
DailyMotion video (s6)	7	98.50 %
E-commerce (w2)	7	98.74 %
iTunes (s2)	7	99.26 %
Facebook video (s3)	7	99.38 %
Generic web (w6)	7	99.51 %
NetFlix (s6)	8	98.97 %
Encrypted web (w5)	8	99.38 %
Twitter (n2)	8	99.40 %
Apple Store (c1)	8	99.43 %
VoIP (x3)	9	95.25 %
Google Drive (c4)	9	98.98 %
Generic cloud (c5)	9	99.29 %
Google Play (c2)	9	99.39 %
Snapchat (m1)	9	99.44 %
Supercell (g4)	9	99.47 %
Generic gaming (g6)	9	96.63 %
Gameloft (g2)	10	85.15%
Mail (m3)	10	99.37%
Adult (w4)	11	98.67%
P2P (x2)	15	96.31%
Gaming platforms (g5)	17	88.83 %
Audio streaming (s8)	17	98.72 %
King (g1)	17	98.85 %
Updates (x1)	18	96.72 %
Pokemon Go (g3)	19	99.09 %
Total number of components		326
Average retained power		98.24 %

Table 4.1
Minimum number of components retaining at least 99% of the total signal power.

A detailed view of the retained components is provided in Table 4.2 for the specific case of the YouTube median week demand. The consumption of YouTube follows four main periodicities in time, namely every day, half-day, 8 and approximately 5 hours. The daily pattern, which is in fact determined by the circadian rhythm of human activities, has a clearly higher impact than the other dynamics in this case.

Figure 4.5 offers an intuitive illustration of the quality of the component filtering process above. Each plot refers to one representative service, and reports: (*i*) the original median week demand (blue) and the inverse DFT (gold) computed using only the components in Table 4.1; and, (*ii*) the residual traffic that is not captured by the inverse DFT (red). In all cases, the inverse DFT allows reproducing the main

Component number	Power	Phase (degrees)	Period
1	7.13%	126.81	24h
2	1.51%	-132.97	12h
3	0.29%	43.95	4.8h
4	0.17%	-85.98	8h

Table 4.2
Example of retained components for the YouTube service.

temporal fluctuations of the original demand, and residuals are limited to low-volume noise. Remarkably, such a good approximation of the traffic time series is obtained with just a few components per service, as detailed in Table 4.1.

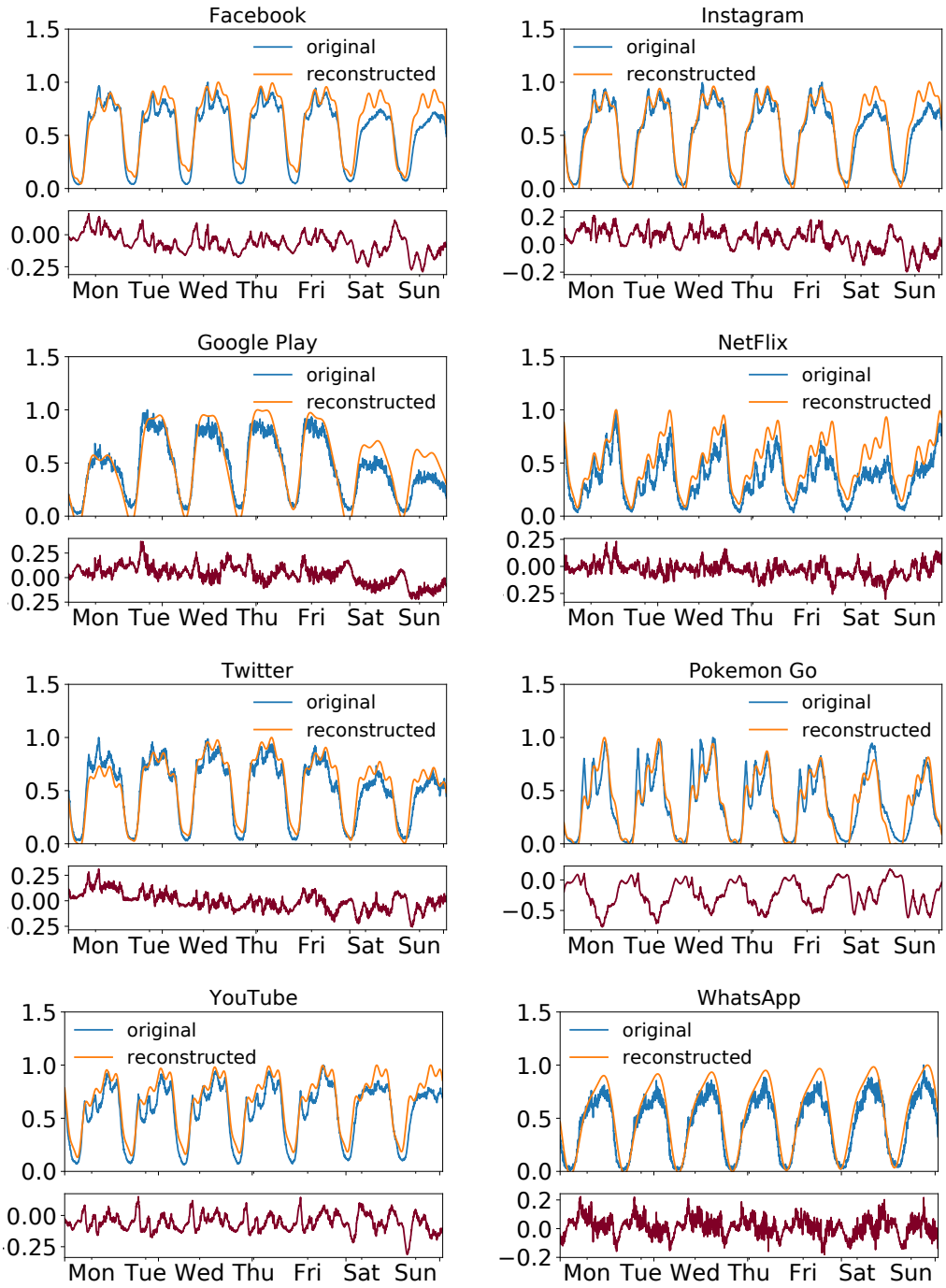


Figure 4.5. Reconstructed traffic demands via the inverse DFT on retained components in Table 4.1.

4.4 Component analysis

By looking at the power spectra in Figure 4.3, it is apparent that many peaks are common to different services. This hints at the fact that the time dynamics of the demands for diverse services may yield patterns that recur at similar periodicity. Next, we design a systematic approach to explore this phenomenon, by developing a simple but efficient clustering algorithm for DFT components in Section 4.4.1, and applying it to our reference service demands in Section 4.4.2.

4.4.1 Clustering components

As explained in Section 4.3.1, each component is characterized by a frequency, an amplitude, and a phase. We implicitly used the amplitude to filter relevant components above, since the power $|X_k^s|^2$ associated to each component is proportional to the square of its amplitude. However, the amplitude is not relevant to the clustering problem; indeed, the amplitude is a measure of the magnitude of a given repetitive pattern, while our objective is to identify similar temporal periodicities across service demands, independently of their magnitude. In other words, if two services feature regular traffic peaks at, *e.g.*, noon every day, we would like to cluster together the components responsible for the peaks, no matter whether they have dissimilar amplitudes because the two services generate different traffic volumes.

Therefore, our clustering algorithm considers only the frequency and phase of each component. The two attributes are in fact processed in two separate steps: in a first step, we group components that have identical frequencies; in the second step, components in the same group are further clustered based on their phase. The rationale for this design is that even slightly different frequencies lead to an increasing misalignment of the components in time, no matter what their phases are: during one week, misaligned components can determine peaks at very different times, which should not be assimilated in our analysis. As a result, we do not want to cluster components with non-identical frequency, even if they have phases that perfectly match. The constraint on equal frequency makes a clustering based on the joint frequency and phase inappropriate, and let us favor a simpler two-step approach instead.

In the first step, we cluster components on their frequency. For the reasons explained above, we require that only identical frequencies are grouped together. Therefore, the clustering operation is straightforward, and we simply gather components based on frequency identity.

The second step focuses on components within each frequency category. In this case, phases that are close but not perfectly matching may be clustered together, since recurring patterns with the same periodicity and small constant shifts in time capture semantically equivalent activity peaks during the whole observation time. Also, the distance measure for phases should be maximum in opposition of phase (*i.e.*, when the value difference is $K\pi, K \in \mathbb{Z}$), and null for phases that are $K2\pi$ apart, $K \in \mathbb{Z}$. To fulfill these specifications, we first map phases to a Cartesian plan; let us denote by ϕ_k^s the phase of the k-th DFT component for the demand of service s, then the Cartesian coordinates are

$$\begin{aligned} x &= \cos \phi_k^s \\ y &= \sin \phi_k^s. \end{aligned} \tag{4}$$

The transformation above places components along a circle of unity radius, at an angle that it is proportional to their phase.

We then run Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a well-known density-based clustering algorithm [68], on the bi-dimensional points that represent the components. We parametrize the algorithm so that at least 3 points shall be grouped to form a cluster, and the maximum distance between the two closest points in the same cluster is $\epsilon = 0.1$, which maps to a phase difference of roughly 5°.

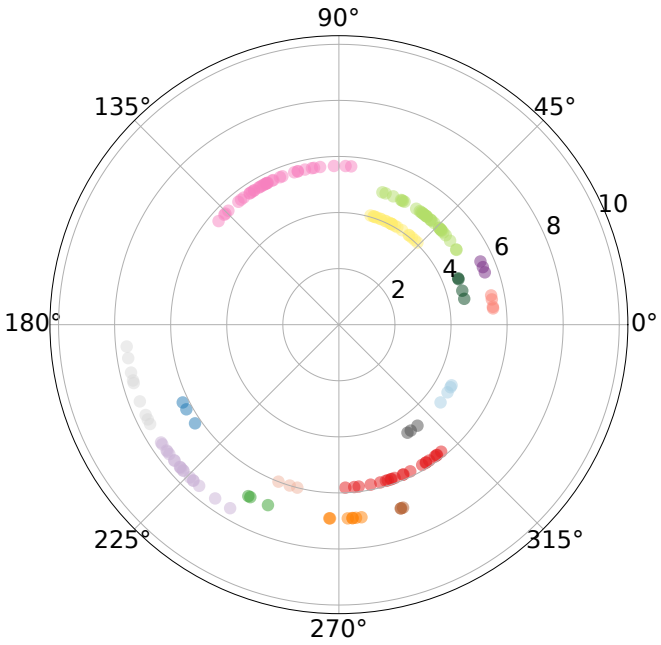
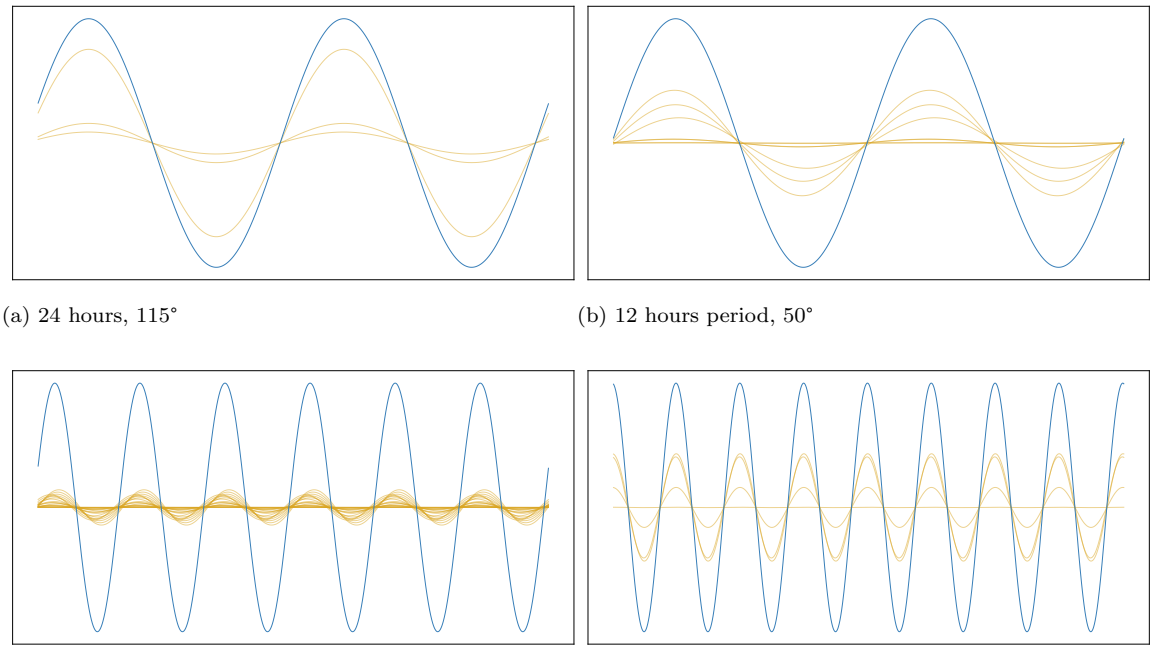


Figure 4.6. Clusters of the 326 components from the demands for the 37 service considered in our study.

4.4.2 Commonalities and outliers in mobile service demands

Figure 4.6 illustrates the 16 clusters obtained with the two-step algorithm above. In order to show the quality of the result jointly for the frequency and phase attributes, we map the frequency to the distance from the origin, and the phase to the angle as per (4). Colors denote points, *i.e.*, components, labelled in the same cluster. This representation outlines clear groups of points, which are well identified by the algorithm, which thus assimilates components with the same frequency (*i.e.*, along the same circle) and close phases as desired.

Figure 4.7 provides a complementary view of the same clustering result. The four plots represent four different clusters, whose period (inverse of frequency) and phase (in °) are indicated below each image. In each plot, every component belonging to the cluster is represented as a sinusoidal function of time (in gold). We can observe that the sinusoids in the same cluster are very similar, hence they correspond to equivalent temporal patterns of activity peaks. The components differ in terms of amplitude, but, as previously mentioned, this is due to the heterogeneous popularity and traffic volume associated with each service. What is relevant to our analysis is the agreement in frequency and phase, which is confirmed by the regular pattern of the sum of components (in blue).



(a) 24 hours, 115°

(b) 12 hours period, 50°

(c) 4.8 hours period, 62°

(d) 8 hours period, 17°

Figure 4.7. Components in four different clusters, portrayed as sinusoidal functions of time.

A comprehensive summary of the 16 clusters is provided in Table 4.3. Columns in the Table correspond to clusters, identified by their frequency and phase indicated in the first row. Subsequent rows refer to each of the 37 services we consider in our analysis. The number in each element (x, y) indicates the fraction of power associated with the component of service x that is sorted in cluster y. The fractional power is calculated with respect to all service components retained for analysis, according to Table 4.1; a value zero in element (x, y) thus indicates that the demand for service x does not have any relevant component with the frequency and phase associated to cluster y. The last column refers to outlier components that could not be associated to any cluster. The last two rows list the number of components in each cluster and the percent power of the whole cluster with respect to all clusters. Table 4.3 lets us provide the following insights.

I. Almost all (33 out 37) of services have a largely dominant component with a 24-hour periodicity. It is easy to map such a component to the circadian rhythm of human activities, which alternates low traffic overnight and high demand during the day.

II. Most (32 out of 37) services also show the same significant dynamic at a 12-hour periodicity. Many (22 out of 37) also share components that highlight regular patterns at every one week, 4.8 hours. An investigation of the causes for these sub-daily patterns is out of the scope here, and an object for future research; yet, we speculate that commuting affects the demands for many services and may be behind these dynamics.

III. Common regular behaviors are present also at periods longer than one day for many (18-21 out of 37) services. One week and 28 hours are the most relevant periods, and we consider that those are linked with different dynamics occurring during weekends.

Cluster period Cluster phase	24h 115°	12h 50°	1w -137°	4.8h 62°	1w -162°	28h -67°	84h -87°	8h 17°	21h 8°	21h 21°	8h -31°	84h -71°	8h -55°	84h -115°	33.6h -107°	42h -150°	none
Ads	84.3	9.2	3.4	2.3	0	0.9	0	0	0	0	0	0	0	0	0	0	0
Adult	19.9	0	12.0	0	0	0	0	0	0	0	0	2.5	0	0	0	0	65.6
Apple Store	61.4	11.6	0	1.0	22.0	1.8	0	0	0	0	0	0	0	0	0	0	2.2
Audio streaming	54.6	0	0	4.9	2.8	1.5	0	0	0	0	0	0	0	1.3	1.1	0	33.8
DailyMotion video	62.2	1.3	22.7		4.0	0	0	4.1	0	0	0	0	0	0	0	0	1.3
E-commerce	65.3	25.5	2.4	3.0	0	0	1.1	0	0	0	0	0	0	0	0	0	2.8
Encrypted web	76.0	9.6	8.3	2.4	0	1.9	0	0	1.1	0	0	0	0.9	0	0	0	0
Facebook	82.1	13.0	1.8	2.1	0	1.0	0	0	0	0	0	0	0	0	0	0	0
Facebook video	79.4	13.3	0	2.7	2.5	0	1.2	0	0	0	1.0	0	0	0	0	0	0
Gameloft	88.7	2.0	0	1.1	0	0	0	0	0	0	0	0	0	0	0	0	8.2
Gaming platforms	80.7	4.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14.7
Generic cloud	76.4	9.7	7.5	0	0	1.5	1.7	0	0	1.0	0	0	1.3	0	0.9	0	0
Generic gaming	86.8	3.5	0	1.5	3.6	1.6	0	0	0	0	0	0	0	0.9	0	0.9	1.3
Generic messaging	96.1	2.3	0	0	1.0	0.6	0	0	0	0	0	0	0	0	0	0	0
Generic video	66.6	22.3	4.3	3.2	0	0	0	3.6	0	0	0	0	0	0	0	0	0
Generic web	83.0	8.5	4.8	1.9	0	1.0	0	0	0	0	0.8	0	0	0	0	0	0
Google Drive	80.3	10.1	0	0	2.6	1.8	0	0	0	0	0	0	0	1.1	1.2	0	2.8
Google Play	61.8	8.5	0	0	20.4	3.6	0	0	1.2	0	0	0	1.4	0	0	0	3.1
Google+	93.3	4.1	0	0.8	0	0.8	0	0	0	0	0	0	0	0	0	0	1.0
Instagram	88.1	9.1	0.8	1.9	0	0	0	0	0	0	0	0	0	0	0	0	0
Instagram video	82.2	14.7	0	1.6	0	0	0	0	0	0	0	0	0	0	0	0	1.6
King	57.8	0	9.1	3.5	0	0	8.2	0	0	0	0	0	0	0	0	0	21.3
MMS	95.1	3.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.6
Mail	66.1	1.3	15.6	0	0	5.1	3.8	0	3.3	0	0	0	0	0	0	0	4.8
NetFlix	0	17.3	0	4.7	0	0	0	2.1	0	0	0	0	0	0	0	0	75.9
News	73.7	13.6	6.6	4.8	0	1.4	0	0	0	0	0	0	0	0	0	0	0
P2P	36.3	0	0	0	0	0	0	0	0	0	0	1.3	0	0	0	0	62.4
Pokemon Go	79.0	0	1.5	4.5	0	0.5	0	0	0	0	0	0	0	0	0	0	14.5
Snapchat	88.7	3.6	0	0	0	1.8	0	0	0	0	0	0	0	0	0	0.9	5.0
Supercell	81.8	3.6	6.3	0	0	2.2	2.3	0	0	1.5	0	0	0	0	0	0	2.3
Twitter	76.5	11.4	6.19	2.1	0	1.5	0	0	0	0	1.4	0.8	0	0	0	0	0
Updates	50.8	10.6	16.9	0	0	2.8	0	0	0	2.9	0	0	0	0	0	0.7	15.4
VoIP	78.7	9.5	5.9	0	0	1.4	1.5	0	0.7	0	0	0	0	0	0	0	2.3
WhatsApp	93.6	6.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
YouTube	78.3	16.6	0	3.2	0	0	0	0	0	0	1.9	0	0	0	0	0	0
iCloud	88.7	7.6	0	0	2.1	1.7	0	0	0	0	0	0	0	0	0	0	0
iTunes	0	31.7	3.1	3.5	0	0	0	7.1	0	0	0	0	0	0	0	0	54.6
Total components	35	32	18	22	9	21	7	4	4	3	4	3	3	3	3	3	115
Percent power	70.7	9.2	3.1	1.7	1.6	1.0	0.5	0.5	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10.8

Table 4.3
Overview of the 16 clusters grouping the 326 retained service demand components.

IV. Several services tend to defy classification, and (*i*) have no or much less relevant components in the 24-hour cluster, (*ii*) have an unusually high weight associated with specific clusters, and/or (*iii*) have a high incidence of components that are not included in any cluster (*i.e.*, are outliers). Services in this category include games (King, Pokemon Go, generic Gaming platforms), audio streaming services (iTunes, generic Audio streaming), NetFlix, peer-to-peer, and adult web traffic. These are fairly specific categories of mobile applications, with different but reasonable reasons for their diversity. For instance, NetFlix is a fairly unique service providing long-lived video streams to niche mobile users. Audio streaming applications are the sole that do not need visual attention by the user. Or, adult web traffic is characterized by unique patterns due to its socially inconvenient nature.

5

Hybrid time-frequency analysis: a deeper view with wavelets

Contents

The spatio-temporal analysis performed in Section 3.3 unveiled the mobile service heterogeneity, while the frequency domain analysis performed in Section 4 allows for a very compact representation of each service time series (*i.e.*, with less than ten components we can almost perfectly reconstruct the original service demand).

However, by using a pure frequency domain analysis, we are not able to properly capture fine-grained metrics such as the variation of the fundamental frequencies along a given period of time. For instance, this occurs when trying to understand weekend - weekdays patterns or variation of the night - day behaviour on subsequent days. Such an effect could be achieved by considering shorter time intervals for the Fast Fourier Transform (FFT) analysis (*e.g.*, one day), but then the selection of the analyzed time fraction may introduce biases in the analysis (*e.g.*, which day is chosen).

We describe in this chapter a hybrid time-frequency analysis that hinges the two proposed approaches to provide a comprehensive view of the services' behaviour, taking into account the intrinsic restrictions of the time - frequency duality (*i.e.*, by observing the frequency only, one cannot evince time variations and vice-versa). We divide the chapter as follows. First, we describe the new analysis methodology in Section 5.1. Then, we show the characterization of data (see Section 5.2), the clusterization of service demands (see Section 5.3) and the spatial variability given this approach (see Section 5.4). Finally, we present the applicability of the insights found with this hybrid time-frequency analysis in Section 5.5

5.1 Flexible analysis combining time and frequency via wavelets

As we are interested in understanding the details of the frequency of the signal in a specific moment in time, we focused in the Wavelet Transformation (WT) analysis as the next step in our study. We look for hidden behavioural patterns inside our data set and the WT provides the flexibility required to slice our signal and analyze its behaviour at the same time both in the time and frequency domain.

It is possible to consider the WT a signal decomposition of the original time series signal onto a group of basis functions [69]. Those basis functions are called wavelets, and are the equivalent of the sinusoidal signals in the frequency analysis with the Fourier transform. The main difference of WT with respect to FFT is the fact that, while the sine waves used in the Fourier analysis have an infinite support (thus they do not discriminate on the temporal dimension), wavelets have a finite response on time and a bounded energy. Thus they are a viable tool to understand time dynamics on the frequency domain.

As wavelets have a finite time domain support, the concept of frequency is replaced by the concept of scale. This measure relates to the way each wavelet is dilated, contracted and shifted during the WT computation. Thus, it can be seen as a representation of the "instantaneous frequency" of each service demand time series.

Being the definition of Wavelets less stringent than the sine waves used in FFTs, there are several families of wavelets available, as thoroughly discussed in [70]. In our analysis we employed the most popular family of wavelet, *i.e.*, the Morlet wavelet, that is defined by:

$$\psi(t) = \pi^{-1/4} e^{iwt} e^{-t^2/2} \tag{5}$$

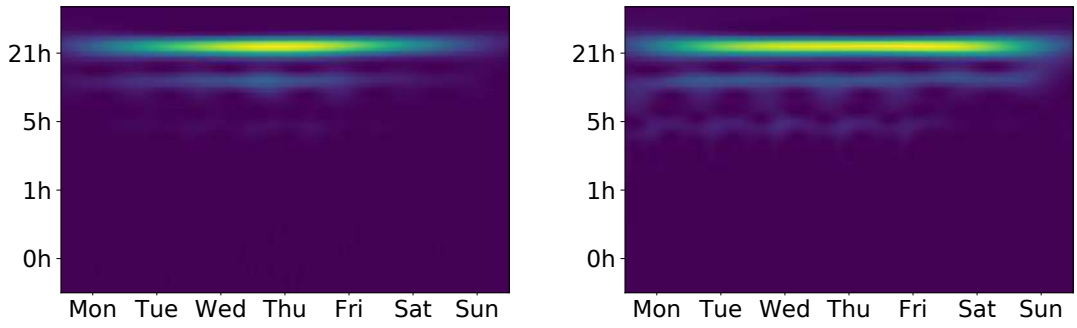
When applying a WT to a time domain signal such as the services demands considered in Chapter 4, the output is a scaleogram: a bidimensional (*i.e.*, time and scale) view of the originally single dimensional data. An example scaleogram is depicted in Figure 5.1, where we show the demand of three services in a period of 7 days. Since our samples are captured every five minutes, a total of 2016 samples are available.

Before applying the WT, the service demand is de-trended (*i.e.*, we subtract the mean and divide by the standard deviation) to avoid any possible distortion in our output. The time resolution for the transformation is configured to 1 sample by default, and the frequency resolution to 20 voices per octave.

Once applied the WT, we obtain the scaleogram of the signal (as represented in Figure 5.1). The x-axis thus shows the time domain information, while on the vertical axis, the scale has been transformed to represent the periods detected in the signal in hours.

5.2 Wavelet scaleogram characterization

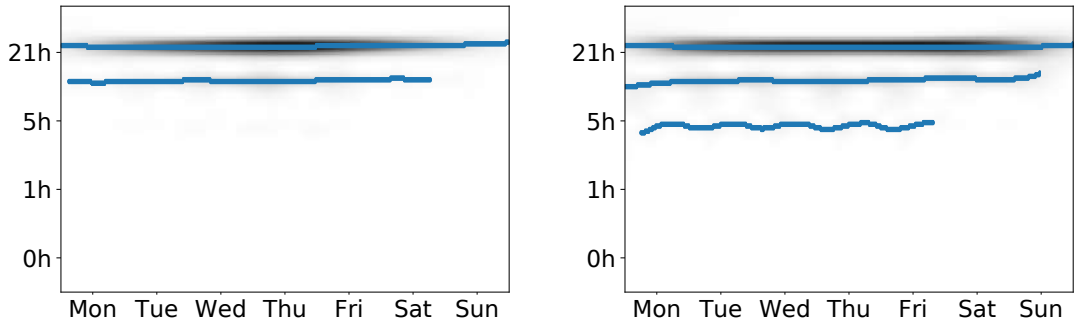
In this section we discuss the similarity assessment across services using the WT based on the comparison of the generated scaleograms. More specifically, we perform a characterization of the scaleograms by identifying their ridges (similarly to the methodology we employed in Section 3.3, in which we were detecting peaks).



(a) Apple Store (b) YouTube

Figure 5.1. Power Scaleogram

A sample scaleogram with ridges looks like the ones in Figure 5.2, and we follow the methodology originally introduced in [71] and subsequently implemented in the WaveletComp R package [72]. For the scope of our research, we ported and adapted the R code [1] to Python.



(a) Apple Store (b) YouTube

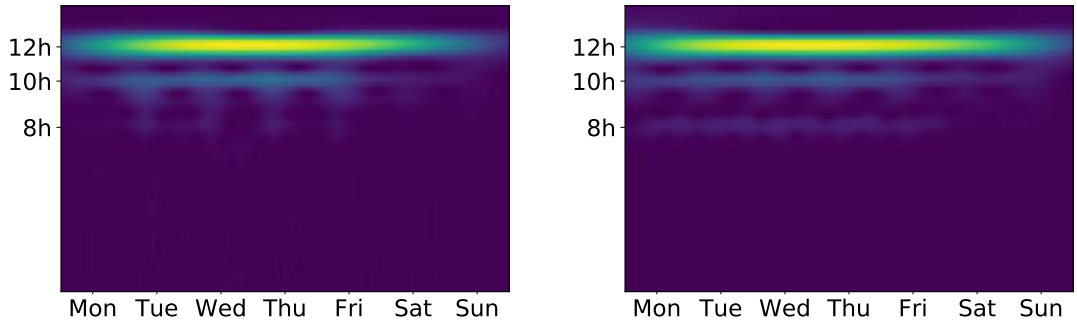
Figure 5.2. Scalogram with power ridge

As previously discussed, the power scaleogram contains the "signature" of a service demand along time, but due to its very large support, it does not provide a compact representation of the time series as, *e.g.*, we could achieve with the FFT. When we do a visual inspection of some signatures, we realize that some services exhibit a similar pattern that could result in a potential clusterization. In Figure 5.3 we can see that some services like Google Play and Generic Web, that were not clusterizable in the time domain, seem to be grouped together in the same slice when we take into account the hybrid time-frequency approach. Thus, we intend to characterize the service demand time series by their power ridges. To this end, we apply the following algorithm.

Our algorithm analyzes all scale values registered for all the times to find the local maxima. If the value is above a predefined percentage θ (in our case 0.05) of the power peak value, a ridge is detected and stored. This computation is done for each time sample, thus generating a new matrix with the ridges of the signal that can be overlaid over the "original" one.

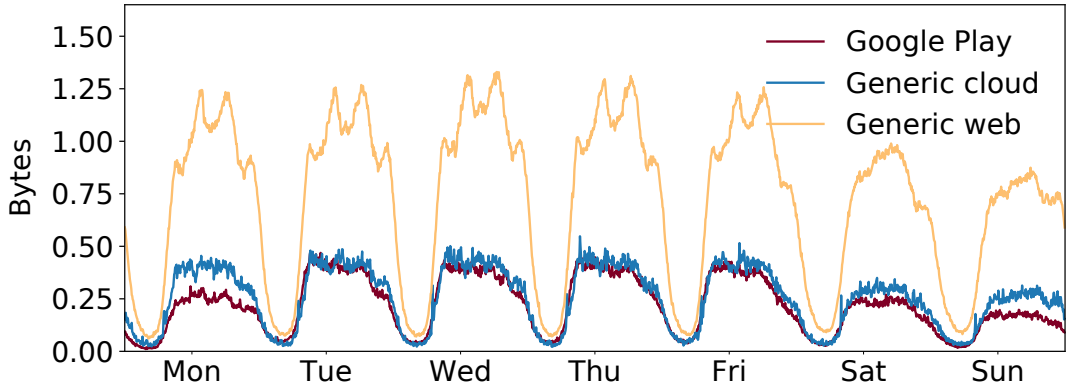
By applying this algorithm we can extract the salient features of each service demand, compressing though the representation of a time varying service demand

[1] The code is available at https://www.github.com/wnluc3m



(a) Google Play

(b) Generic web

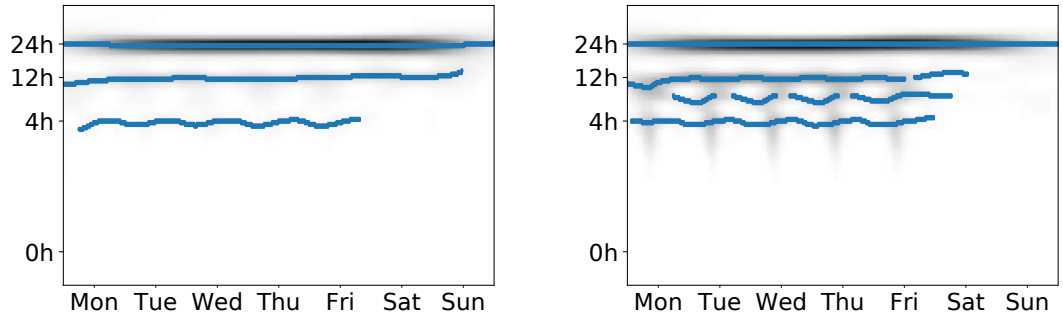


(c) Service demand

Figure 5.3. Wavelet vs Service demand

into a binary matrix with ones representing the location in the time-scale domain of a power ridge.

As shown in Figure 5.4 for a selected number of services, we can identify consistent usage patterns around several periods of 4h, 8h, 12h, and 24h. We remark that this characterization is more flexible than the one used in the FFT representation, which has to resort to a very compact, yet coarser, representation of the different components.



(a) YouTube

(b) Audio streaming

Figure 5.4. Scalogram with power ridge

5.3 Ridge-based service demand clustering

Despite being quite compact (*i.e.*, a binary matrix), the dimensionality of the ridge representation is still high. Hence, in order to study commonalities across services, we further compress the resulting ridges matrix to feed a clustering algorithm.

Our approach consists of a split of the ridges representation into several partitions of size τ along the time domain. The target is to summarize each partition with the relevant scales that feature a ridge, reducing thus the number of input data from several thousands values per range (*i.e.*, the full scale support), to tens of values.

As an example, we show the output of this algorithm for the ridge representation of YouTube discussed before, by selecting partitions of $\tau = 24$ hours (see Table 5.1). We can see a clear dichotomy between weekdays (that exhibit ridges for scales of 8h, 10h, and 12h), Saturday's ridges at 10h, and 12h, up to Sundays, that can be condensed into just two ridges components at 12h. With this compact yet flexible representation, we proceed to cluster the service demand time series looking for similarities.

Service	Ridge 1	Ridge 2	Ridge 3
Monday	8h	10h	12h
Tuesday	8h	10h	12h
Wednesday	8h	10h	12h
Thursday	8h	10h	12h
Friday	8h	10h	12h
Saturday	10h	12h	-
Sunday	12h	-	-

Table 5.1
Ridge found on Figure 5.2 with τ =24h

Similarly to the methodology employed for the FFT analysis, we hence define a similarity metric between services according to the placement of their ridges along the week. Again, as also performed in Chapter 4, instead of defining an ad-hoc similarity metric, we simply use euclidean distances computed on a transformation of the data points extracted from the ridge sampling, as detailed in Table 5.1.

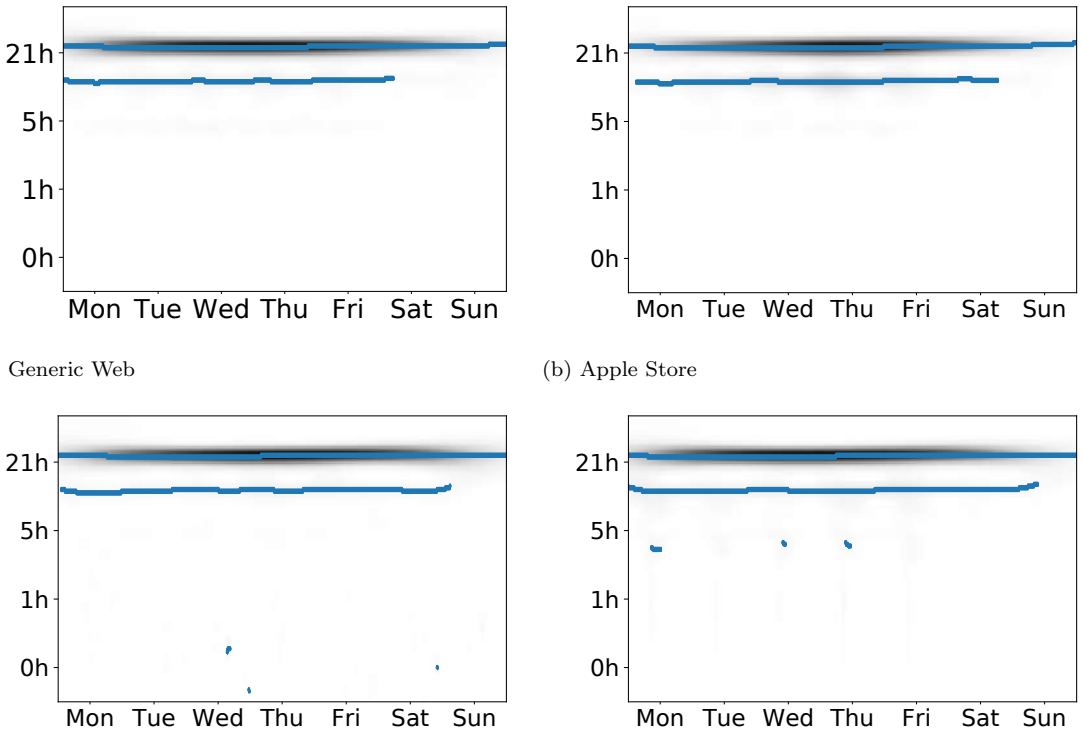
Thus, we condensate the n-dimensional feature vector of each service (where the features are the ridge locations, juxtaposed day by day) by applying a Principal Component Analysis (PCA) with number of dimensions equals to 2. Besides retaining most of the variability of each service (the explained variance ratio is equal to 0.83), having just two dimensions allows for a powerful representation of each service into a two dimensional plane.

Also in this case, we use DBSCAN as a clustering algorithm with default parameters (`min_samples = 2` , `eps = 2`), obtaining 4 clusters plus an additional one for outliers, as detailed in Table 5.2.

From Table 5.2, it is clear that clustering 37 distinct service demands is possible in a few clusters, but the result is not optimum. As we can observe in Figure 5.5, there are some services that have similar ridge behavior, but some others could be clustered much better. The reasoning behind this "sub-optimal result" is that we are trying to cluster together services that represent 37 dimensions into only 2 dimensions. Hence, in the next section we explore an alternative clustering approach which considers how close base stations are in the city under study. To this end, we divide it into 16 areas and consider a similar volume of traffic.

Cluster	Services
1	Supercell, P2P
2	Generic video, Audio streaming, VoIP
3	Facebook video, NetFlix, DailyMotion Google Play, Generic messaging, Generic gaming
4	E-commerce, Generic web, Apple Store, iCloud Google Drive, Instagram, Twitter, Updates, Pokemon Go
Outliers	YouTube, iTunes, Instagram video, Facebook Google Services, Ads, News, Adult, Encrypted web Generic cloud, King, Gameloft, Gaming platforms Snapchat, WhatsApp, Mail, MMS

Table 5.2
Clusterization of ridges found like on Figure 5.2 with τ =24h



(a) Generic Web

(b) Apple Store

(c) iCloud

(d) Google Drive

Figure 5.5. Example of services' ridges in Cluster 4 from table 5.5.

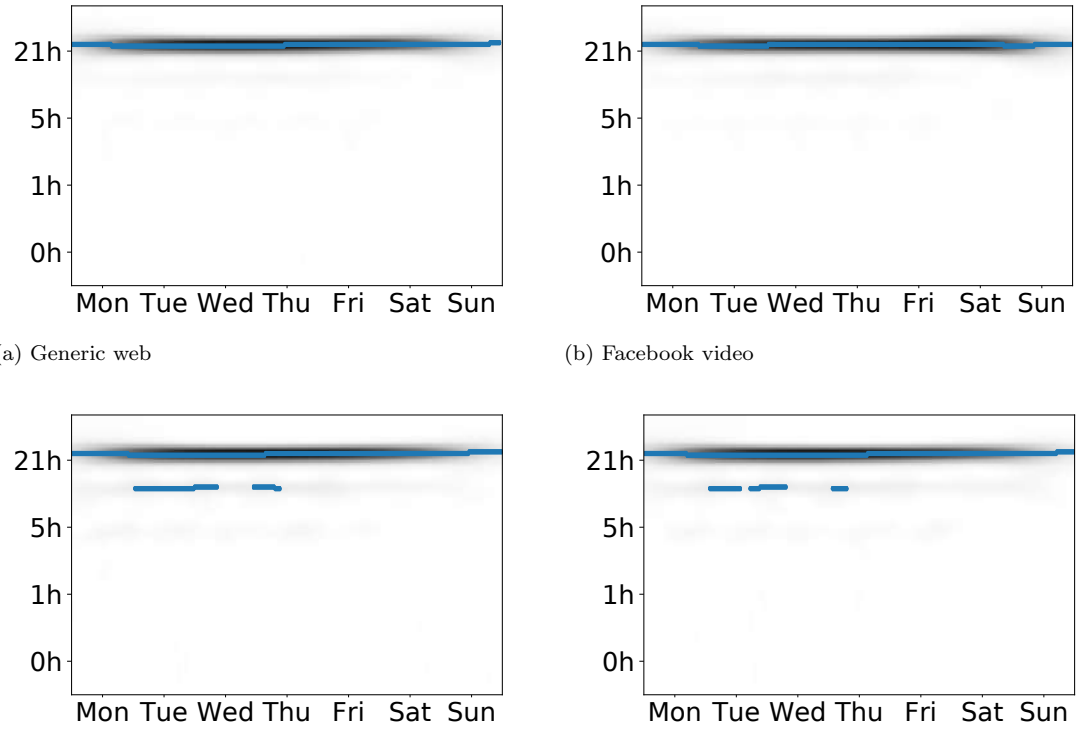
5.4 Spatial variability of temporal analysis

As in [3], our evaluation we assume that our area under study could be divided in hierarchical network structures. Since we are studying the reference scenario of the metropolis, the operator is able to deploy a number N_ℓ of nodes at the generic level ℓ, each responsible for a subset of the antenna sites at the radio access level. We assume the same methodology as in [3], where the operator deploys generic level-ℓ nodes and links based on two criteria: (i) the offered load shall be similar at all nodes; (ii) the subset of antennas served by the same node shall be geographically contiguous. Jointly,

these criteria represent a plausible strategy that aims at maximizing the performance of network slicing, ensuring load balancing and reducing latency between antenna sites and nodes.

In this section, we focus on $\ell = 8$ (*i.e.*, $N_\ell = 16$), allocating a static (and arbitrarily set) amount of resources. However, our model can easily accommodate different definitions of the number of nodes determining the load balance, which would apply to specific network hierarchies. We remark that a static allocation of resources may be acceptable in the current 4G monolithic architecture, but will become extremely expensive across the high number of network slices expected to characterize next-generation systems.

From Table 5.3, we can see that in most of the defined regions, it makes sense to have at least two clusters for distinct service's behaviors. We illustrate the case in the first region. This region presents two sub-clusters of services, meaning that we can find two different temporal patterns between applications in the same area. As we can observe in Figure 5.6, the services clearly exhibit two distinct behaviors inside the same cluster, each of them correctly clustered. Besides, if we compare the ridges of Generic Web that come from the first region (the ones from Figure 5.5) versus the Generic Web ridges version of the full city in Figure 5.5, we can also see that the ridges vary. This means that each service could behave differently depending on the area.



(a) Generic web

(b) Facebook video

(c) Generic gaming

(d) Gameloft

Figure 5.6. Similarities of ridges per region in Cluster 1.1 (top) and Cluster 1.2 (bottom) from Table 5.3.

Region	Clustered services (wavelet fingerprint)
1	(w6) (w5) (w3) (s3) (s4) (m4) (m1) (m3) (g4) (g3) (n1) (n2) (n3) (n4)
1	(w4) (g6) (g2) (x2) (x3)
2	(w6) (m4) (m1) (g4)
2	(w4) (g6) (g5) (g2) (x3)
2	(s3) (n1) (n2)
3	(w1) (s3) (n1) (n2)
3	(m4) (m1) (g4) (n3)
3	(g6) (g5) (g2) (x3)
4	(w6) (w5) (w1) (s4) (m5) (m4) (m1) (g4) (n2) (n3) (n4)
4	(g6) (g5) (g2) (x3) (x1)
5	(w6) (w5) (w1)
5	(s3) (s1) (s4) (n1) (n2) (n4)
5	(g2) (x2) (x3)
6	(w6) (w1) (n4)
6	(g6) (g5) (g2) (x3)
7	(w3) (w1) (s3) (s4) (c4) (n1) (n2)
7	(m4) (m1) (g4) (n3)
8	(w6) (m4) (m3) (g4) (n3)
8	(w5) (m1) (n2) (n4)
8	(w4) (g5) (g2) (x3)
9	(w4) (g6) (g5) (g2) (x3)
9	(w1) (n1) (n2)
9	(m1) (m3) (g4)
10	(w6) (w1) (s3) (s4) (m4) (m1) (c4) (n1) (n2) (n3) (n4)
10	(w5) (w3) (m3) (g4)
10	(g6) (g5) (g2) (x3)
11	(s1) (s4) (n1) (n4)
11	(c3) (g2) (x2) (x3) (x1)
12	(w6) (w5) (w3) (w2) (w1) (s3) (s1) (s4) (m4) (m1) (c4) (n1) (n2) (n3) (n4)
12	(w4) (g6) (g5) (g2) (x2)
12	(m3) (g4) (c1)
13	(w6) (w1) (n1)
13	(m1) (g4) (n2) (n3) (n4)
13	(g6) (g5) (g2) (x3)
14	(w6) (w5) (w1) (s3) (s1) (s4) (n1) (n2) (n4)
14	(w4) (g6) (g5) (g2) (x3) (x1)
15	(w4) (g6) (g5) (g2) (x3)
16	(g6) (g5) (g2) (x2) (x3)

Table 5.3
Similarities of ridges per region. Service shortening is described in Table 4.1.

5.5 Applications to composability

As the accelerated digital era will result in a gigantic increase of mobile data traffic, it is of paramount importance to develop distinct techniques that could be adopted by network operators to handle this load within the context of 5G networks. Since network slicing is one of the preferred techniques that would allow them to orchestrate their complex network structures, it is crucial to manage resources efficiently.

Our first spatio-temporal approach (see Chapter 3) revealed that service requirements are heterogeneous, and the clustering based on it would not be the best approach. However, our next data-driven clustering results are significantly meaningful for network orchestration and planning purposes, as the operator decides the number of antennas that are required to process all the data and the region where a data center is needed. The proposed techniques, complementary to the definition of network slices and hierarchies in the related work, provide some light in regards to take informed decisions. We not only fill the gap on current possible clustering methodologies with reasonable performance, but also show the trade-off between complexity and efficiency in data's computational effort. We emphasize that for the static scenarios, the hybrid approach implies a reduction of the data with PCA, in order to achieve a high-speed performance. However, for the frequency approach the task that consumes more effort resides in tuning the parameters to have a reasonable number of discarded components without biasing the result.

Moreover, we should mention that such data reductions when implementing slices will be key in dynamic 5G scenarios, where a quick reconfiguration of the network is expected. In particular, wavelets have shown the best applicability for understanding service requirements geographically. The decomposition of service's traffic data in several patterns, exhibited distinct behaviors even for a specific service. This fact allows the operator to define clusters based on similar patterns and less time periods than the spectral approach, aiming to complement the one-side spectral proposal that defined many additional clusters.

Hence, the trade-off between the complexity of the defined virtual network structure and the number of slices depends upon a new variable: the number of regions where a service has a similar usage pattern. For instance, we can observe that YouTube has up to 8 different patterns, and therefore the network operator could tune the network slice not only considering the number of clusters of a city, but also taking into account the number of regions to cover (see left graph in Figure 5.7).

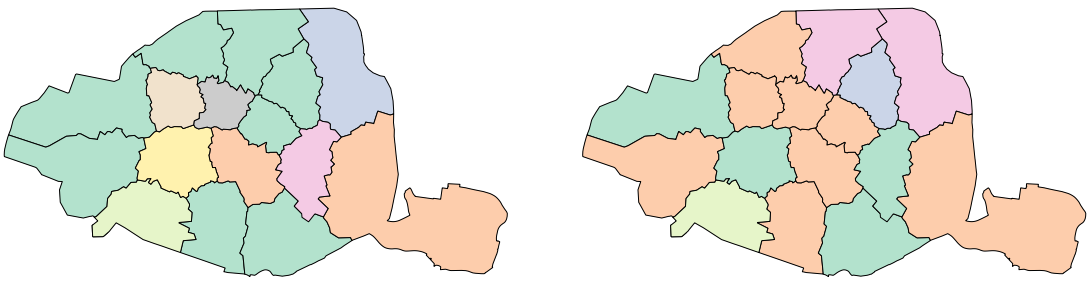


Figure 5.7. Similarities of ridges per region for YouTube (left) and Instagram (right).

In fact, the wavelet signature of a service can be clusterized over space given its own patterns, and this is illustrated in Figure 5.7. In these graphs, we can clearly identify that both YouTube and Instagram services have analogous areas defined for their hybrid time-frequency characterization. This means that an operator could decide

to cluster together both services (*i.e.*, group them in the same customized network slice) over those regions that share the same color for orchestration purposes, as long as their network requirements are compatible.

Under this scheme, services that are clusterized together tend to use more resources during the same period of time, as they are synchronized, whereas the non-clusterized services potentially do not suffer this effect. While we leave the orchestration algorithm for future work, in the next section we quantify the efficiency of putting together different services in the cloud.

Last but not least, we claim that the novel methodology to generate and work with synthetic data solves data privacy derived issues stated in the GDPR. This technique, combined with the wavelets approach and forecasting algorithms would allow operators to design their network infrastructure in advance, reducing the over-provisioning risk and boosting new networks based on distinct data traffic consumption.

6

Data-driven resource management

Contents

Multi-service networks [20] are a key building block for the implementation of the network slicing paradigm [26] that, in turn, will enable new business models such as multi-tenancy [40] and finally pave the way to 5G. At this stage, the bulk of the work on next generation network sharing architectures is already available, ranging from novel visions of the network [73] to specific architectures proposals [74]. More specifically, research work already addressed the extension to multi-service settings of

fundamental parts of the 5G system, such as the RAN [75,76], the core network [77], or the management and orchestration components [78]. Such research effort is already making its way into standardization: 3GPP considered multi-service and network slicing aspects for the next Release 15 [79].

On top of the architectural research work, enabling multi-service networks has also been considered from an algorithmic point of view. The focal point of research in this area has been RAN resource allocation [39, 80, 81], as oversubscribing spectrum is especially difficult. However, resource sharing has also been tackled for other kinds of virtualized functions [32].

Despite the attention that multi-service networks, network slicing and multi-tenant networks have been receiving for the last few years, little attention has been paid to how such network slices will behave in practical scenarios. Understanding the system efficiency *in the wild* has only been possible in reduced scenarios involving very few devices [39], or by making assumption on the real patterns, modelling user movements and service requests with random processes [82]. The only works that employ a data source comparable to ours are the one in [83] and our seminal work in [4].

Our work sheds light on this overlooked aspect in this chapter, and it is organized as follows. First, we introduce an empirical evaluation of slicing efficiency for large-scale scenarios in Section 6.1. Second, we describe the slice specification in 6.2. Third, we define two possible methods for resource allocation in terms of time slot and traffic volume in Section 6.3, followed by the multiplexing efficiency in presence of realistic multi-service demands (see Section 6.4). Next, we discuss our reference scenarios in Section 6.5, where we perform several data-driven evaluations (see Section 6.6). Finally, we summarize all the conclusions extracted from the distinct experiments in Section 6.7.

6.1 Network slicing scenario and metrics

In this section we expose our network model, as well as our representation of the slice QoS requirements and their associated resource allocation strategy. We also introduce the metrics we adopt to evaluate the resource sharing performance.

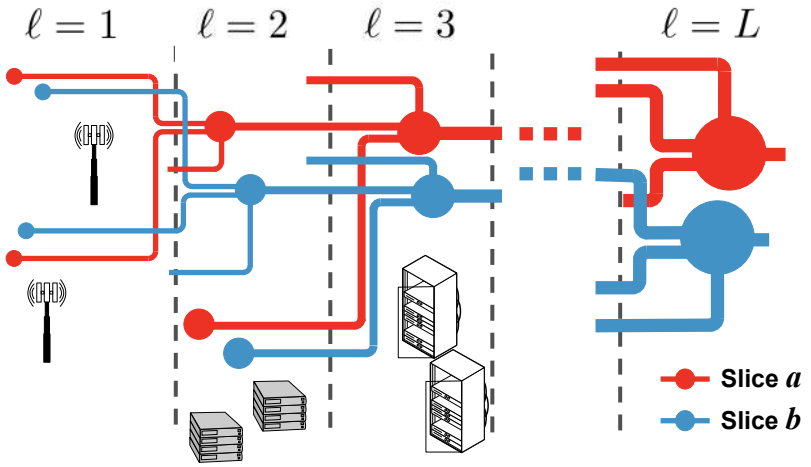


Figure 6.1. Mobile network architecture. The mobile traffic in each slice (*e.g.*, a or b) is increasingly aggregated as it flows from radio access to network core.

Let us consider a mobile network providing coverage to a geographical region where mobile subscribers consume a variety of services. The network operator implements slices $s \in \mathcal{S}$, each dedicated to a different subset of services.

We assume that each slice can be implemented according to any of the strategies in Figure 2.1. To capture such a general scenario, we model the mobile network architecture as a hierarchy composed by a fixed number of levels ($\ell = 1, \ldots, L$) ordered from the most distributed ($\ell = 1$) to the most centralized ($\ell = L$), as illustrated in Figure 6.1. Every network level ℓ is composed by a set $\mathcal{C}_\ell$ of network nodes, each serving a given number of base stations. In the two extremes, we have $\ell = 1$, where network nodes in $\mathcal{C}_1$ have a bijective mapping to individual antennas, and $\ell = L$, where $\mathcal{C}_L$ contains a single network node controlling all antennas in the whole target region. In between, for $1 < \ell < L$, the number of network nodes in $\mathcal{C}_\ell$ decreases with ℓ, whereas that of base stations served by each such node increases accordingly. Note that, in general, a node $c \in \mathcal{C}_\ell$ will operate on data flows that are increasingly aggregated with ℓ, which, as we will see, has a significant impact on resource management.

This hierarchical representation allows considering a variety of node types, along with their associated (possibly virtual) network functions. At the most distributed level ($\ell = 1$), each node runs functions that operate at the antenna level, *e.g.*, concern spectrum or airtime resources. In intermediate cases ($1 < \ell < L$), nodes are at first in charge of a small number of antenna sites, *e.g.*, C-RAN datacenters running VNFs such as dedicated baseband processing or radio resource management. As ℓ grows, VNFs are pushed further towards the network core, into telco-cloud datacenters that tunnel traffic to and from large sets of antenna sites: there, VNFs customize VM resources for large traffic volumes associated to the services delivered by each tenant to subscribers in wide geographical areas. In the limit case ($\ell = L$), all traffic in the target region is managed in a fully-centralized fashion at a single datacenter.

Ultimately, the layered network model allows generalizing our analysis to diverse VNFs, by studying the system performance at different network levels. This also implicitly accommodates all of the network slicing strategies outlined in Figure 2.1. Slices of *type-D* and *type-E* deal with the lowest network layers that are implemented at the antennas, hence correspond to $\ell = 1$. Slices of *type-A* refer to VNFs operating at higher network layers that are deployed at centralized cloud datacenters, hence correspond to high values of the network level ℓ. Slices of *type-B* and *type-C* are concerned with VNFs at radio access, *i.e.*, at base stations ($\ell = 1$), or at higher architectural levels ($1 < \ell < L$) in a C-RAN implementation.

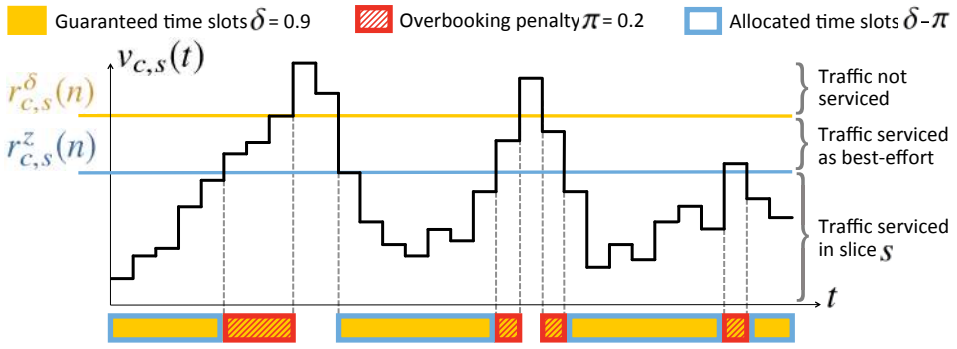
Note that we do not require that a single network deploys virtualization technologies at all network levels. Instead, by taking a large number of levels and considering each of them in isolation, this approach lets us cover a wide range of deployment options and provide insights for all of them.

6.2 Slice specifications

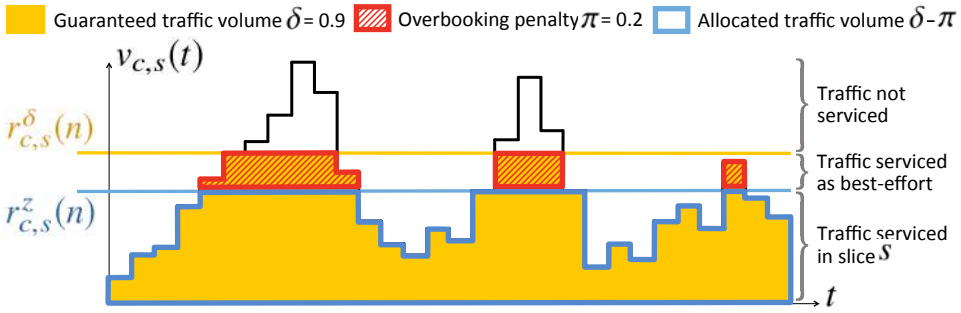
Network slicing primarily aims at letting the operator fulfill the QoS requirements requested by each tenant. To model such requirements, we consider discrete-time demands associated to slices, by averaging traffic over *time slots* denoted by t. Let $v_{c,s}(t)$ be the traffic demand associated to slice s at node c during slot t, as in Figure 6.2. We capture the QoS requirements of s as a *slice specification* z defined by two features.

6.2.1 Guaranteed demand δ

The operator engages to guarantee that the total traffic demand of the slice is fully serviced for a portion at least $\delta \in [0, 1]$, which can be expressed in terms of time or traffic. In the first case, the operator assures that the slice demand is fulfilled during a fraction δ of time slots, as in Figure 6.2a. In the second case, the slice demand is serviced for a fraction at least δ of its volume, as in Figure 6.2b.



(a) Time slot δ



(b) Traffic volume δ

Figure 6.2. Example of resource allocation to a slice s at node c, under guaranteed demand $\delta = 0.9$ and overbooking penalty $\pi = 0.2$, during one reconfiguration period n.

6.2.2 Overbooking penalty π

The operator can decide to overbook network resources to multiple slices, transparently to the tenants [84]. Similar to common practices in the airline or hotel industries, this management model allocates the same resources to multiple tenants, expecting that some will ultimately not use all of their booked capacity; if this is not the case, and services actually require all of the reserved capacity, overbooking leads to violations of the guaranteed demand δ.

Through overbooking, the operator can maximize its revenues by properly balancing the cost of allocated resources and the penalty associated with violations [84]. In our model, we do not adopt a specific overbooking strategy; instead, we consider that the strategy selected by the operator produces violations for a portion $\pi \leq \delta$ of the total traffic demand. This implies that only a fraction of traffic $\delta - \pi$ is actually serviced by the slice, while the portion π of violated demand is treated as best-effort

traffic by the operator. This approach can capture any overbooking strategy, and lets us investigate how violations of δ affect savings in allocated resources. We remark that π may be a fraction of time slots or a fraction of traffic volume, consistently with the representation of δ: the two situations are illustrated in Figure 6.2a and Figure 6.2b, respectively.

In the first case, the slice specification is expressed in terms of time slots, hence the discrete-time traffic of the slice, $v_{c,s}(t)$, is serviced for 90% of the time slots, denoted by the filled (yellow) temporal interval below the abscissa. Due to overbooking, demand δ is violated in 20% of the total time slots, highlighted by the (red) pattern intervals below the abscissa.

In the second case, the slice specification is expressed in terms of traffic, hence $v_{c,s}(t)$ is serviced for 90% of its volume, denoted by the filled (yellow) area under the time series. Due to overbooking, demand δ is violated for 20% of the total volume, highlighted by the (red) pattern region.

6.3 Resource allocation to one slice

We denote a slice specification characterized by a guaranteed demand δ and an overbooking penalty π as $z = (\delta, \pi)$, which becomes more stringent for higher values of δ and smaller π. The operator shall then ensure that enough resources are dedicated to the slice so as to meet z. We now expound the expression of the resources allocated to a slice s by the mobile network operator under a generic $z = (\delta, \pi)$.

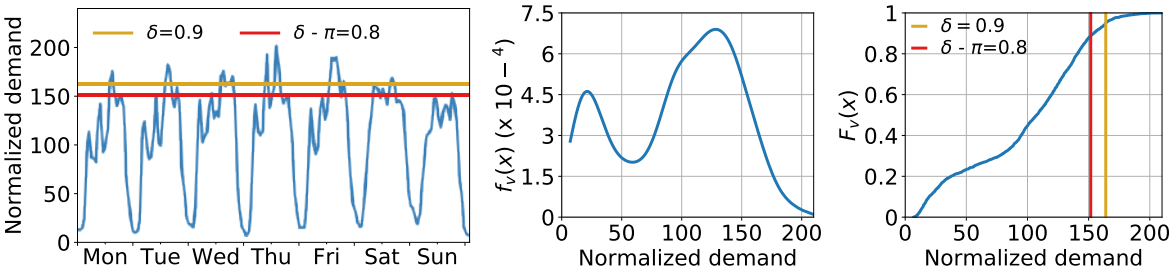
In presence of algorithms that enable a dynamic reconfiguration of VNFs, the resource allocation can be re-modulated over time. In practice, however, the periodicity of reconfiguration is limited by the technological constraints of the slicing strategy adopted (see Figure 2.1). For instance, when network slicing is performed at the antenna level, times in the order of minutes are needed to turn on and off the radio-frequency front-end and reset the transport network. When dealing with radio resource management algorithms (*i.e.*, dynamic spectrum or multi-provider scheduling), re-assignments are constrained by signalling overhead. Or, in the case of VM orchestration, the timescale is limited by instantiation and migration delays [85].

Let us assume that τ is the minimum amount of time steps needed for resource reallocation, which we refer to as a *reconfiguration period*. We denote by $n \in \mathcal{T}$ the n^{th} reconfiguration period within the set $\mathcal{T}$ of all reconfiguration periods that compose the system observation time; n can be then seen as the set of τ time steps it encompasses, *i.e.*, $n = \{t, \dots, t + \tau - 1\}$. During period n, we name $r^{\delta}_{c,s}(n)$ the minimum amount of resources that allow meeting the guaranteed demand δ for slice s at node c. Equivalently, $r^{z}_{c,s}(n)$ is the amount of resources that fulfill z, accounting for both δ and the overbooking penalty π. The formalism is the same when δ is a fraction of time or traffic, as shown in Figure 6.2. Then, our objective is the computation of $r^{z}_{c,s}(n)$, which represents the resources actually allocated by the operator to slice s at node c, based on $v_{c,s}(t)$ and z. Since calculations are different depending on whether δ (hence π) is expressed in terms of time or traffic, below we discuss these two instances separately. For the sake of readability, in the following we drop the c, s, and n notation, and refer to a generic slice, network node, and reconfiguration interval; hence $v(t)$ and r^{z} stand for $v_{c,s}(t)$ and $r^{z}_{c,s}(n)$, respectively.

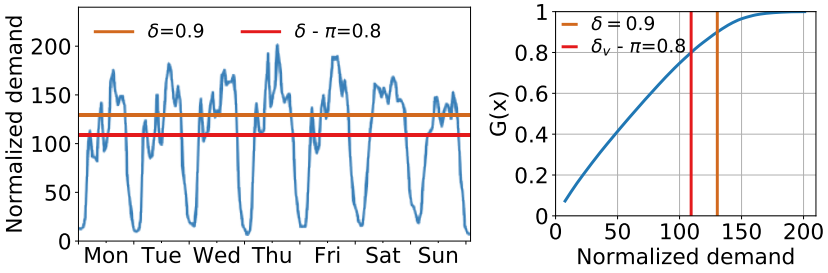
6.3.1 Time slot fraction δ

In this case, the allocation of resources in the target reconfiguration period is such that the offered load $v(t)$ exceeds r^z for a fraction $\delta - \pi$ of the time slots in the reconfiguration period, as shown in Figure 6.2a. This can be formalized as $P\big(v(t) \leq r^z\big) = \delta - \pi$, $\forall t$, where $P(\cdot)$ denotes the probability of the argument. Let f_v be the Probability Density Function (PDF) of the demand, *i.e.*, $f_v(x) = P\big(v(t) = x\big)$. Then, the CDF of the demand $v(t)$ in the reconfiguration period is $F_v(x) = \sum_{y=0}^{x} f_v(y) = P\big(v(t) \leq x\big)$. Therefore, the original condition above is $F_v\big(r^z\big) = \delta - \pi$, and the minimum r satisfying the actual guaranteed demand is $r^z = F_v^{-1}(\delta - \pi)$.

Figure 6.3a illustrates this concept in a practical example. Top left graph shows the weekly time series of the mobile traffic demand for a slice s at a network node c. The horizontal lines denote the minimum resources r^δ and r^z to be allocated when $\tau = 1$ week. Top middle graph (a) shows a representation of $f_v(x)$. In addition, we present $F_v(x)$ (left graph) with cuts at δ and $\delta - \pi$ that identify the needed resource r^δ and r^z, respectively [1].



(a) Time slot



(b) Traffic volume

Figure 6.3. Example of resource allocation to a slice with specification $z = (\delta, \pi) = (0.9, 0.1)$

.

6.3.2 Traffic volume fraction δ

When the operator guarantees (and overbooks) a fraction of traffic, we do not reason in time slots anymore, but account for the effective demand volume associated to each time slot. For this purpose, we introduce a water-filling function, that computes the overall fraction of served traffic as a function of the assigned resources r. Specifically, we define $G(x) = \sum_t \left(\min\left(v(t), x\right)\right) / \sum_t v(t)$, for all time slots t in the target

[1] Traffic volumes in Figure 6.3 as well as in the rest of the result reported in the thesis are normalized with respect to the minimum average traffic recorded at a 4G antenna sector in our reference scenarios presented in Section 6.5.

reconfiguration period. Through the above expression of $G(x) \in [0,1]$, the value of x maps to the upper limit of a water-filling algorithm. The minimum r^z satisfying the actual guaranteed demand is then $r^z = G_v^{-1}(\delta - \pi)$.

Figure 6.3b illustrates this concept in a practical case. Left graph (b) shows the weekly time series of the mobile traffic demand for a slice s at a network node c. As aforementioned, the horizontal lines denote the minimum resources r^δ and r^z to be allocated when $\tau = 1$ week. Besides, we describe $G(x)$ on the right plot (b) with cuts at δ and $\delta - \pi$ that identify the needed resource r^δ and r^z, respectively.

Note that in both cases above, the expressions of r assume that the amount resources needed to serve a given slice is directly proportional to the mobile traffic demand in that slice. While this clearly holds for some types of resources (*e.g.*, radio), we acknowledge that it may be a strong simplification in other settings. We argue, however, that it is a reasonable assumption for many practical VNFs. Moreover, this choice allows us to investigate through a unified framework different network levels ℓ, where resources map to diverse physical assets (such as spectrum, airtime, Central Processing Unit (CPU) time, computational power, or memory) depending on ℓ.

6.4 Multiplexing efficiency definition

Having computed $r^z_{c,s}(n)$ according to either model in Section 6.3, we can define the amount of dedicated resources that the operator allocates to network slices at network level ℓ, over the entire system observation period, as

$$\mathbb{D}^z_{\ell,\tau} = \sum_{s\in\mathcal{S}} \sum_{c\in\mathcal{C}_\ell} \sum_{n\in\mathcal{T}} \tau \cdot r^z_{c,s}(n). \tag{6}$$

Equation (6) covers the demand that receives dedicated resources within slices. However, under overbooking, a fraction π of traffic is penalized, *i.e.*, is treated as best-effort. Such traffic is not isolated anymore, and can be aggregated into a single time series described, at node c and during period n, as

$$v_c(t) = \sum_{s\in\mathcal{S}} \max\left\{0, \min\left\{r^\delta_{c,s}(n), v_{c,s}(t)\right\} - r^z_{c,s}(n)\right\}.\ t \in n, \tag{7}$$

This equation computes the penalized traffic in a slice s as the difference between the resources dedicated to the slice, $r^z_{c,s}(n)$, and those that would be actually needed to accommodate the guaranteed demand, $r^\delta_{c,s}(n)$. As exemplified in Figures 6.3a and 6.3b, $r^\delta_{c,s}(n)$ can be computed by the strategy in Section 6.3, as $F_v^{-1}(\delta)$ or $G_v^{-1}(\delta)$, for the cases where δ is a fraction of time slots or traffic volume, respectively. Then, the shared resources required to serve all traffic penalized by overbooking are, trivially, $r_c(n) = \max_{t\in n} v_c(t)$. Finally, we can calculate the total amount of resources that the operator needs to allocate at network level ℓ, in order to meet specifications z, as

$$\mathbb{R}^z_{\ell,\tau} = \mathbb{D}^z_{\ell,\tau} + \sum_{c\in\mathcal{C}_\ell} \sum_{n\in\mathcal{T}} \tau \cdot r_c(n). \tag{8}$$

Equation (8) returns the total amount of resources that the operator needs to provision at network level ℓ in order to satisfy its commitments with all tenants, when dynamically re-configuring[2] the allocation with periodicity τ, and according to its

[2]Equation (8) maps to the special case where no reconfiguration is possible at level ℓ, when τ is the total system observation time, *i.e.*, $|\mathcal{T}| = 1$.

designated overbooking strategy. In order to unveil the implications of this value, we compare it against a *perfect sharing* benchmark. In perfect sharing, the allocated resources correspond to those required when there is no isolation among different services, hence traffic multiplexing is maximum.

Let $u_c(t) = \sum_{s \in \mathcal{S}} v_{c,s}(t)$ be the total demand for mobile data traffic at node c, summed over all slices. We then denote by $\hat{r}_c^{\delta}(n)$ the resources needed to accommodate $u_c(t)$ during reconfiguration period n. For the sake of fairness, the same requirement δ on guaranteed demand is enforced here as well[3]. Thus, adopting the methodology presented in Section 6.3, $\hat{r}_c^{\delta}(n)$ can be computed as $F_u^{-1}(\delta)$ or $G_u^{-1}(\delta)$, where $F_u(x)$ and $G_u(x)$ are the CDF of the total demand $u(t),\ t \in n$, expressed in time slots and traffic volume, respectively. The resources allocated under perfect sharing are then computed as

$$\mathbb{P}_{\ell,\tau}^{\delta} = \sum_{c \in \mathcal{C}_\ell} \sum_{n \in \mathcal{T}} \tau \cdot \hat{r}_c^{\delta}(n). \tag{9}$$

Taking the above benchmark, we define the *multiplexing efficiency* as the ratio between the resources required with perfect sharing and those needed under network slicing, *i.e.*,

$$\mathbb{E}_{\ell,\tau}^{z} = \mathbb{P}_{\ell,\tau}^{\delta} / \ \mathbb{R}_{\ell,\tau}^{z}. \tag{10}$$

In summary, $\mathbb{E}_{\ell,\tau}^{z}$ quantifies the efficiency of network slicing in terms of resource management at network level ℓ, under resource reconfiguration intervals of duration τ, and with slice specification $z = (\delta, \pi)$. As $\mathbb{E}_{\ell,\tau}^{z}$ approaches one, the total amount of slice-isolated resources tends to that assured by a perfect sharing. As slicing the network becomes increasingly capacity-demanding, the efficiency drops instead towards zero.

Let us illustrate the operation of multiplexing efficiency in Figure 6.4, when δ is expressed as a fraction of time slot (top) or of traffic volume (bottom). The left column depicts the time series of the mobile traffic demand for a set $\mathcal{S}$ of five slices, observed at a single network node c, during one reconfiguration interval n (τ = 1 week). A slice specification $z = (\delta, \pi) = (0.9, 0)$ commits the operator to allocate, for each slice s, at least the capacity marked by the grey horizontal lines, which are computed as discussed in Section 6.3. Their sum, in thick gold, denotes $\sum_{s \in \mathcal{S}} r_{c,s}^{z}(n) + r_c(n)$, *i.e.*, the value that, once multiplied by τ, returns the resources specified by Equation (8), at a single node c and during reconfiguration interval n.

The right column, instead, shows the time series of the traffic demand aggregated over all slices in $\mathcal{S}$. By applying the specification z, we get a value $\hat{r}_c^{z}(n)$, highlighted by the horizontal thick gold line. Its multiplication by τ gives the equivalent capacity needed under *perfect sharing* as per Equation (9). Then, the multiplexing efficiency is the ratio between the values highlighted by the thick gold lines on the right and left plots, respectively. In this toy example, the value on the left is only slightly higher than that on the right, hence $\mathbb{E} \sim 1$ and resource isolation is efficient. This is not necessarily the case in practical scenarios, as we will detail later.

[3]We remark that the notion of overbooking penalty is meaningless under perfect sharing, as all traffic is aggregated and treated as best-effort already.

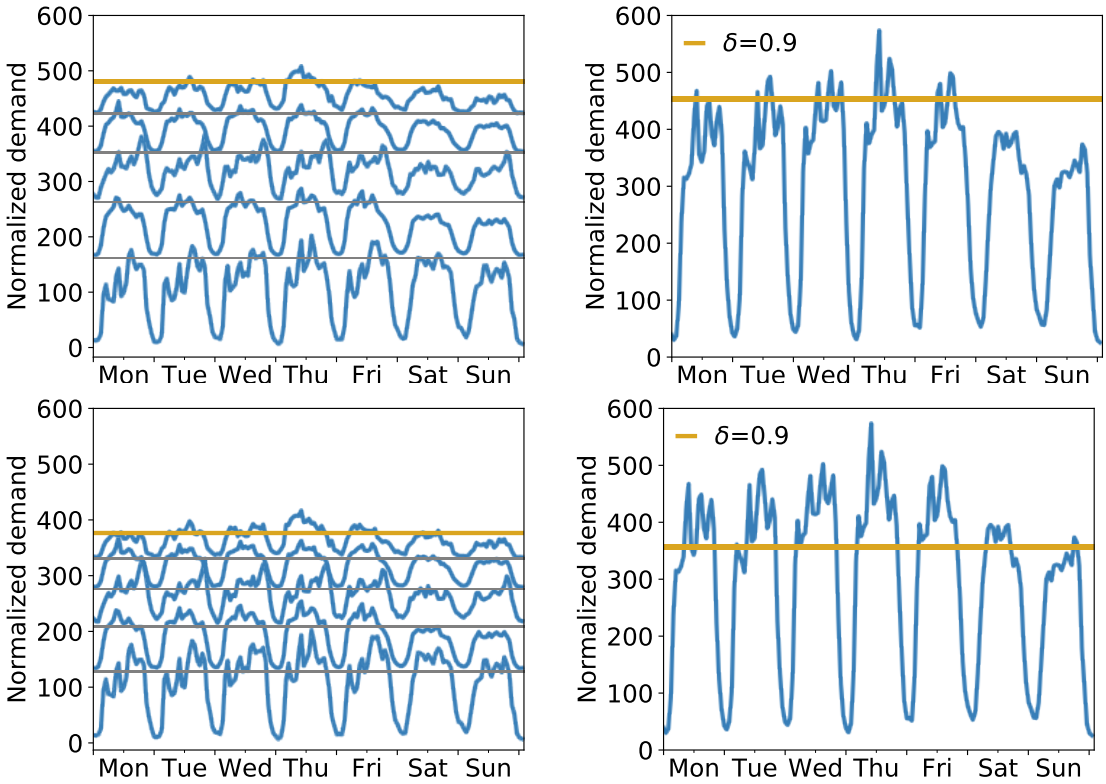


Figure 6.4. Examples of multiplexing efficiency, when $\delta = 0.9$ expressed in time slots (top) and traffic volume (bottom).

6.5 Reference scenarios

In this chapter, we evaluate the efficiency of resource management in a sliced network by considering two modern metropolitan-scale network scenarios. As mentioned in Section 1, today's mobile services already offer a variety of requirements that makes it meaningful to investigate the impact of slice isolation on network efficiency with present traffic.

Our two reference urban regions are the large metropolis of several millions of inhabitants considered in Section 4, and a typical medium-sized city with a population of around 500,000, both situated in Europe. Service-level measurement data was collected in the target areas by Orange. Details are in Section 6.5.1. On top of this, we model the hierarchical network infrastructures in the target regions by assuming a deployment of nodes that balances load and reduces latency. This is discussed in Section 6.5.2.

6.5.1 Mobile service demands

The real-world demands generated by individual mobile services in the two reference regions were collected during three months in late 2016. The information was gathered by the proprietary, as well as aggregated geographically (per antenna sector) and temporally (over 5-minute time intervals), so as to make the data non-personal and to preserve user privacy; all operations were carried out within the operator premises, under control of the local DPO, and in compliance with applicable regulations (see Section 2.4).

The resulting measurement data describe downlink and uplink traffic for hundreds of prominent mobile services consumed in the target regions. Building on such information, we define potential slices by identifying mobile services that meet two requirements: (*i*) they generate a substantial offered load (above 0.1% of the total network traffic), sufficient to justify a dedicated network slice; and (*ii*) they have clear KPIs and QoS requirements. We identify 37 services that meet the criteria above, and associate them to a different network slice each.

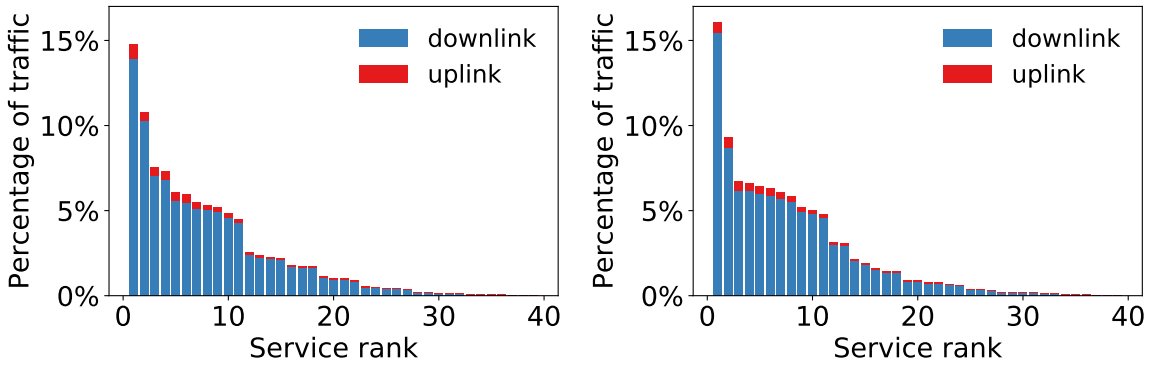


Figure 6.5. Percentage of the mobile traffic generated by the selected services. Different colors denote downlink and uplink traffic. Left: large metropolis. Right: medium-sized city.

Our choice of services represents well the heterogeneous nature of today's mobile traffic. It encompasses many popular services, such as YouTube, NetFlix, Snapchat, Pokemon Go, Facebook or Instagram, and covers a wide range of classes with diverse network requirements, including mobile broadband (*e.g.*, long-lived and short-lived video streaming), low-latency (*e.g.*, gaming, messaging), and best effort (*e.g.*, web browsing, social media), which are representative forerunners of 5G services [86]. Figure 6.5 provides basic information on our selection of services. It outlines the downlink-dominated, highly skewed traffic split among the services, whose percent traffic can differ of more than two orders of magnitude.

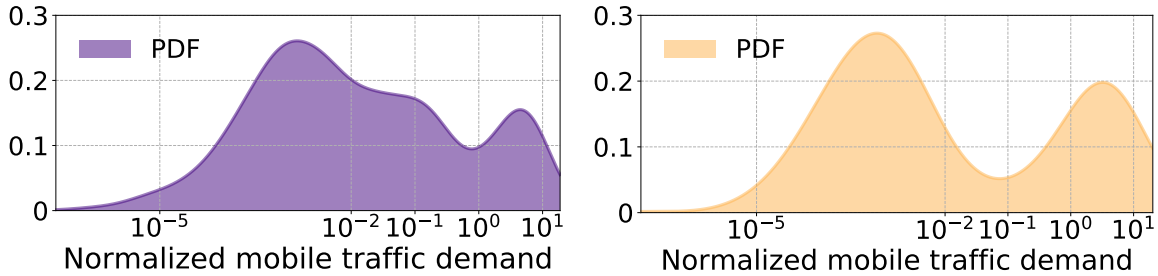


Figure 6.6. PDF of the traffic demands across all antenna sectors. Left: large metropolis. Right: medium-sized city.

A strong diversity also emerges in the way the selected services are consumed across the geographical space within the two urban regions. Figure 6.6 portrays the PDF of the total offered load at individual antenna sectors, which again spans several orders of magnitude. The main cause of heterogeneity is the radio access technology: our measurement data captures 2G, 3G, and 4G access, and 4G antennas accommodate

much larger fractions of the demand and generate the rightmost bell-shaped lob of the distributions. Still, 10-time differences in the traffic volume appear even across 4G antenna sectors, implying substantial location-based demand variability.

6.5.2 Hierarchical network structure

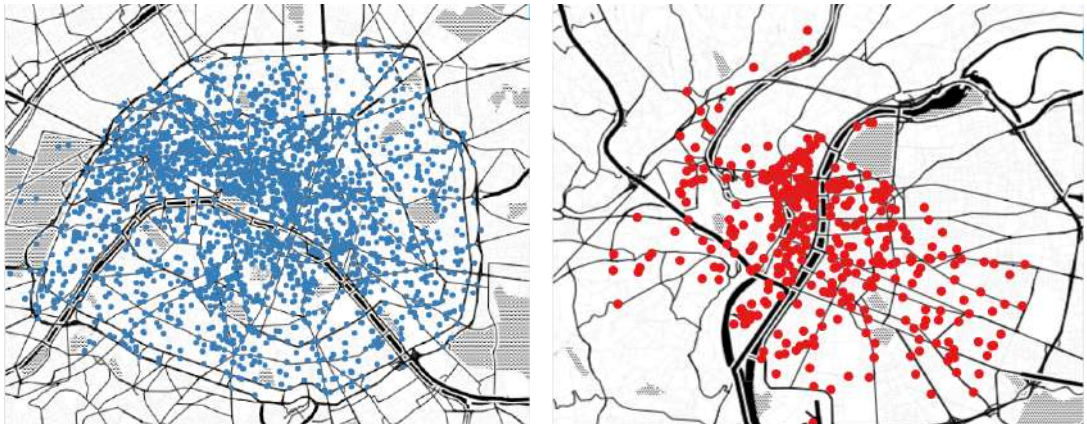


Figure 6.7. Antenna deployments in the target regions. Left: large metropolis. Right: medium-sized city.

The deployment of antennas in the target regions is shown in Figure 6.7, which highlights the diversity of the case studies in terms of network infrastructure, owing to the different geographical span and user population density of the two areas. While we do not have information on the architecture of the mobile networks beyond the radio access, we model the hierarchical structure exemplified in Figure 6.1 after current proposals for cloudified network slicing [73], as follows.

At the generic level ℓ, the operator deploys a number $N_\ell = |\mathcal{C}_\ell|$ of nodes, each responsible for a subset of the antenna sites at the radio access level. Every node will thus run VNFs (whose nature will depend on ℓ) on the mobile data traffic incoming from or outgoing to its associated antennas. We assume that the operator deploys generic level-ℓ nodes and links based on two criteria: (i) the offered load shall be similar at all nodes; (ii) the subset of antennas served by a same node shall be geographically contiguous. The first criterion ensures load balancing, and the second reduces latency between antenna sites and nodes. Jointly, these criteria represent a plausible strategy that aims at maximizing the performance of network slicing. We remark that the resulting node deployment is static and does not change during our experiments; instead, the node resources allocated to each slice may change under dynamic resource allocation schemes.

Under these criteria, the problem of associating the level-ℓ nodes with the original antenna sites in Figure 6.7 is a special case of *balanced graph k-partitioning*. Let us consider a graph where each vertex $v \in V$ maps to one antenna site, and has an associated cost $c(v)$ equal to the mobile traffic demand recorded at the site; also, let an edge $e = \{u, v\} \in E$ connect vertices u and v only if the corresponding antenna sites are geographically adjacent[4]. The problem of level-ℓ node-to-antenna site association

[4]Multiple notions of adjacency are possible. We opt for one that leverages the common practice of approximating antenna coverage areas via a Voronoi tessellation: two sites are then adjacent if they share one Voronoi cell side.

translates into dividing the graph into N_ℓ sub-graphs, such that the sum of costs of nodes in each partition is balanced. We introduce decisions variables

$$e_{uv} = \begin{cases} 1 & \text{if } e \text{ is a cut edge} \\ 0 & \text{otherwise} \end{cases} \qquad \forall e \in E, \tag{11}$$

$$x_{v,k} = \begin{cases} 1 & \text{if } v \text{ is in partition } k \\ 0 & \text{otherwise} \end{cases} \qquad \forall v \in V, \forall k, \tag{12}$$

and formulate an Integer Linear Programming (ILP) problem:

$$\min \sum_{e_{uv} \in E} e_{uv} \tag{13}$$

$$\text{s.t.} \sum_{v \in V} x_{v,k} \cdot c(v) \leq (1+\epsilon) \cdot \frac{\sum_{v \in V} c(v)}{N_\ell}, \ \forall k \tag{14}$$

$$\sum_{v \in V} x_{v,k} \cdot c(v) \geq (1-\epsilon) \cdot \frac{\sum_{v \in V} c(v)}{N_\ell}, \ \forall k \tag{15}$$

$$\sum_{k} x_{v,k} = 1, \ \forall v \in V. \tag{16}$$

$$e_{uv} \geq x_{u,k} - x_{v,k}, \ \forall e \in E, \forall k \tag{17}$$

$$e_{uv} \geq x_{v,k} - x_{u,k}, \ \forall e \in E, \forall k \tag{18}$$

The objective function given by Equation (13) aims at minimizing the number of cut edges that join vertices in separate partitions, so as to generate graph subsets that are as compact as possible. Our goal in terms of load balancing is ensured by the constraints given by Equations (14) and (15), which bound the load difference among the various subsets of antennas: each partition is forced to have a total cost that is within a fraction ϵ from the ideal case of a perfectly even cost $\sum_{v \in V} c(v)/N_\ell$. The constraint given by Equation (16) ensures that each vertex is in exactly one partition, while those given by Equations (17) and (18) determine the value of decision variables e_{uv} based on whether vertices u and v belong to a same partition as defined by $x_{u,k}$ and $x_{v,k}$.

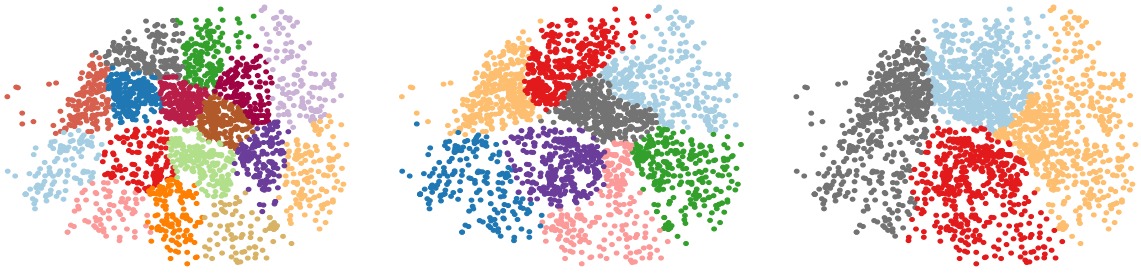


Figure 6.8. Association of antenna sites to level-ℓ nodes in the large metropolis scenario. The plots refer to $\ell = 8$ (16 nodes, left), $\ell = 9$ (8 nodes, middle) and $\ell = 10$ (4 nodes, right).

The resulting optimization problem is NP-hard. We use a suitably configured version of the Karlsruhe Fast Flow Partitioner (KaFFPa) heuristic [87] to solve it,

which previously requires a Delaunay tesselation structure [88] in order to compute the total number of edges and neighbors of the antennas under study given their coordinates. In doing so, we allow for ±10% imbalance among the load served by nodes at every level ℓ, *i.e.*, $\epsilon = 0.1$ in Equations (14) and (15). Figure 6.8 shows three examples of antenna site partitioning among network nodes, for a selection of levels ℓ in the large metropolis scenario. [5]

Table 6.1 summarizes instead the main features of the partitions obtained in our two urban scenarios. Rows are (*i*) the level $\ell \in \{1, \ldots, 12\}$, (*ii*) the corresponding normalized mobile traffic per node, and *(iii)-(iv)* the number of nodes $\mathcal{N}_\ell$ serving each urban region at network level ℓ. At $\ell = 1$, nodes map to 4G antenna sectors, and the traffic per node is an average. From $\ell = 2$ to $\ell = L$, we consider the partitions obtained by solving the optimization problem given by Equation (13), where $L = 12$ for the large metropolis and $L = 10$ for the medium-sized city.

ℓ		1	2	3	4	5	6	7	8	9	10	11	12
Traffic per node		5	10	15	30	60	75	100	150	300	600	1167	2334
$\mathcal{N}_\ell$	Metropolis	422	230	160	80	40	32	23	16	8	4	2	1
	City	122	60	40	20	10	8	6	4	2	1		

Table 6.1
Hierarchical network structures for two urban scenarios.

6.6 Data-driven evaluation

Our performance evaluation is organized as follows. First, we investigate worst-case settings where very stringent slice specifications are enforced and no reconfiguration is possible (Section 6.6.1). We then relax these constraints, and assess efficiency as slice specifications are moderated (Section 6.6.2), as well as under a dynamic orchestration of network resources (Section 6.6.3). We then investigate the impact of a varying number of slices on efficiency (Section 6.6.4). Afterwards, we explore a number of meaningful, specific case studies among all possible system configurations (Section 6.6.5), and finally we discuss the efficiency from a different view (Section 6.6.6).

6.6.1 Slicing efficiency in worst-case settings

The least efficient sliced network scenario implies (*i*) strict slice specifications, where the mobile network operator commits to guarantee the whole traffic demand ($\delta = 1$) for all slices, (*ii*) no possibility of overbooking ($\pi = 0$), and (*iii*) a static allocation of resources without option for reconfiguration over time (τ spans the whole three-month observation time in our measurement dataset, and $|\mathcal{T}_\tau| = 1$). With this configuration, the operator trades efficiency for simplicity: it replicates physical resources for different slices, and statically allocates to each slice the resources needed to meet the associated offered load. This strategy yields the lowest efficiency in terms of occupied physical resources, but does not require any advanced solution for dynamic resource management to be implemented in the network. It could be a pragmatic approach to practical network slicing, if the loss of efficiency is small.

[5] Note that graph partitioning is only used to outline plausible deployments where node load is reasonably balanced, yet, as we do not require a perfect balance, the specific partitioning algorithm is of no particular relevance.

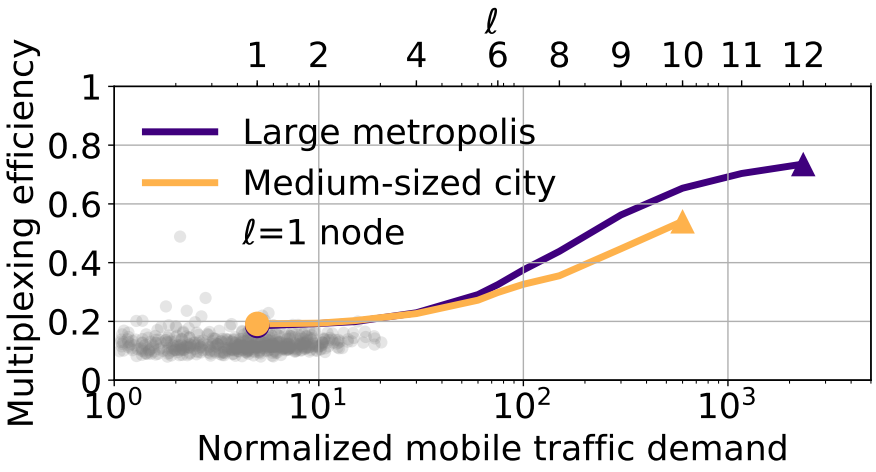


Figure 6.9. Efficiency of slice multiplexing versus the normalized mobile traffic served by one node (bottom x axis) at level ℓ (top x axis) in the two reference urban scenarios.

Figure 6.9 portrays the multiplexing efficiency of slicing as a function of the network hierarchy level ℓ (top x axis); for the sake of clarity, the latter is mapped to the normalized mobile traffic demand observed by a level-ℓ node (bottom x axis), as per Table 6.1. The two curves refer to our two reference urban scenarios, and outline the fluctuation of the efficiency as one moves from resources at the antenna level (dot on the left) to those in a fully centralized cloud (triangle on the right). Results are for the static resource assignment previously described, *i.e.*, $|\mathcal{T}_\tau| = 1$, and slice specification $z = (\delta, \pi) = (1, 0)$. Dots denote $\ell = 1$ and triangles $\ell = L$, for each scenario. Scattered grey points around $\ell = 1$ denote the efficiency and traffic measured at all level-1 nodes (*i.e.*, individual 4G antenna sectors) separately. These results, and all others unless stated otherwise, refer to the case where the 16 mobile services that generate the most network traffic are allocated to independent slices each; the rationale for this choice will be apparent when discussing the effect of a varied number of slices, in Section 6.6.4.

The curves in Figure 6.9 confirm the intuition that the efficiency grows as one moves from very distributed resources at the antenna level to more centralized ones. This trend roots in the temporal dynamics of traffic in the difference slices: the demands for each slice are typically very bursty at individual antenna sectors, whereas aggregating demands over a growing number of base stations results in increasingly smoother time series. The coefficients of variation of the traffic time series substantiate this conjecture: their values range in $[1.487, 2.363]$ for $\ell = 1$ and in $[0.511, 0.587]$ for $\ell = L$, with intermediate levels resulting in midway ranges. The erratic, high activity peaks that occur at the antenna level ($\ell = 1$) force the allocation of substantial static resources in order to accommodate the per-slice traffic. For higher ℓ values, peak-to-average ratios are instead substantially reduced, mitigating these effects and increasing the overall efficiency.

In addition to the qualitative trend with ℓ, Figure 6.9 lets us appreciate the following quantitative results on the efficiency.

- The efficiency is very low ($\sim$0.19) at the antenna level: ensuring physical resource isolation across slices in absence of dynamic reconfiguration capabilities would require more than 5 times the capacity of a legacy architecture where no network slicing is implemented. The grey points highlight that such a poor efficiency affects all 4G antenna sectors, independently of their specific offered load.

- The efficiency grows slowly when aggregating traffic at the network edge ($\ell = 2$ to $\ell = 6$). Some gain starts to be appreciable as one moves above $\ell = 7$ in our reference scenarios, *i.e.*, at network nodes that accommodate the demands from many tens of antenna sectors at least.
- However, in absolute terms, even when considering that all traffic generated in each of our two urban scenarios is aggregated at a single level-L node ($L = \{12, 10\}$ in the large metropolis and medium-sized city, respectively, see Table 6.1), the efficiency stays fairly low, at 0.54–0.74.

We note that, although the method presented in Section 6.3 operates on individual levels separately, Figure 6.9 offers a complete view of end-to-end efficiency across the network, and the result covers all of the different types of slices presented in Section 1. For instance, a *type-A* slicing in Figure 2.1 limits the analysis to the rightmost part of the plot: implementing the most basic form of slicing requires roughly doubling the resources deployed in the network core cloud with respect to a legacy non-sliced case. More complicated slices that reach deeper into the network architecture encompass larger portions of the curves in Figure 6.9. As an example, let us image that a *type-C* slice in Figure 2.1 corresponds to a network level $\ell = 6$ in a specific infrastructure layout: then, the plot details the loss of efficiency that the operator can expect at all intermediate nodes, down to a threefold increase of required resources at the C-RAN datacenters that lie at the very edge of the slice. Furthermore, when considering an end-to-end network slice, we have that the slice can be associated to resources located at different levels of the network infrastructure (e.g., some resources at the antenna $\ell = 1$ and others at the core $\ell = L$). In this case, the resulting overall efficiency of the network slice is the combination of the individual efficiencies of the resources deployed at different levels.

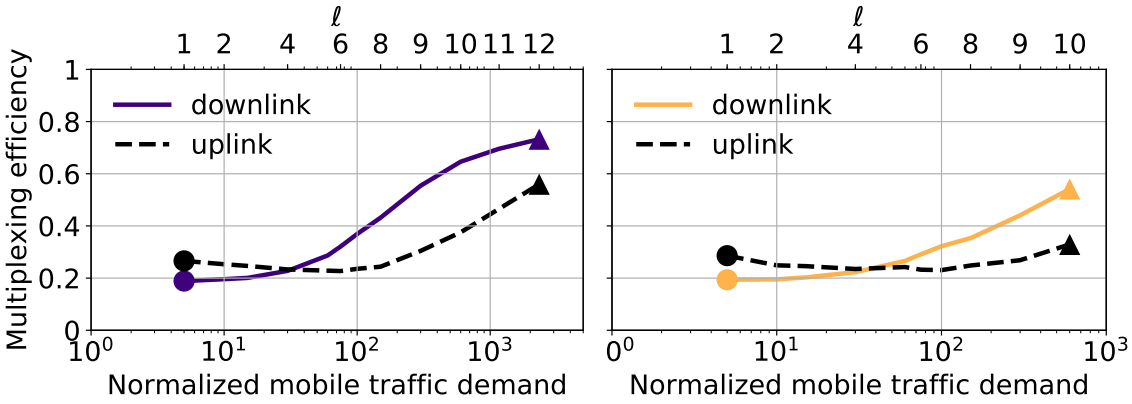


Figure 6.10. Efficiency of slice multiplexing, in the same settings of Figure 6.9, separating downlink and uplink. Left: large metropolis. Right: medium-sized city.

The results can be further disaggregated for the downlink and uplink directions, as shown in Figure 6.10. Downlink traffic dominates the total demand, as previously seen in Figure 6.5: therefore, the associated efficiency curves are very close to those in Figure 6.9. However, the trend of efficiency during uploads is sensibly different from the global one: slicing traffic in uplink tends to become remarkably (40% to 60%) less efficient as one moves towards more centralized network levels. We argue that the reason lies in the small uplink traffic volume, which results in bursty time series with high peak-to-average ratios, even upon aggregation over multiple antennas.

The distinct trends for downlink and uplink are especially important in the light of the different costs associated to the demands in the two directions. By looking at the sheer traffic load, the overall resource assignment should be driven by the downlink behavior, since it currently dominates the aggregate data volumes, as per Figure 6.5. However, specific applications, hence slices, heavily rely on uplink traffic: for instance, the fact that efficiency at the antenna level is also low in uplink means that services with strong requirements on access network latency (*e.g.*, mobile gaming) are as hard to accommodate as downlink bandwidth-eager ones (*e.g.*, video streaming). As another example, baseband processing at a virtualized radio access is remarkably more CPU-intensive for uplink traffic [89]: the very low efficiency recorded in uplink at the network edge can make resources assignment challenging when dealing with *type-C*, *type-D* or *type-E* slices in Figure 2.1.

An interesting final remark on the results in Figures 6.9 and 6.10 is that we do not observe substantial differences between the two reference cities. Minor discrepancies only emerge for high values of ℓ, and can be easily imputed to the intrinsic topological and demographic differences that characterize the two scenarios. The affinity of results in the two different urban regions is in fact a constant across all results, as it will be observed in the remainder of this section.

6.6.2 Configuring slice specifications

Severe slice specifications may represent a major cause for the poor efficiency recorded in Section 6.6.1. To gain insight on this, we investigate the impact that the QoS requirements imposed on each slice have on the opportunities for multiplexing slice demands. Note that here we still consider a static allocation of resources, and no possibility of reconfiguration.

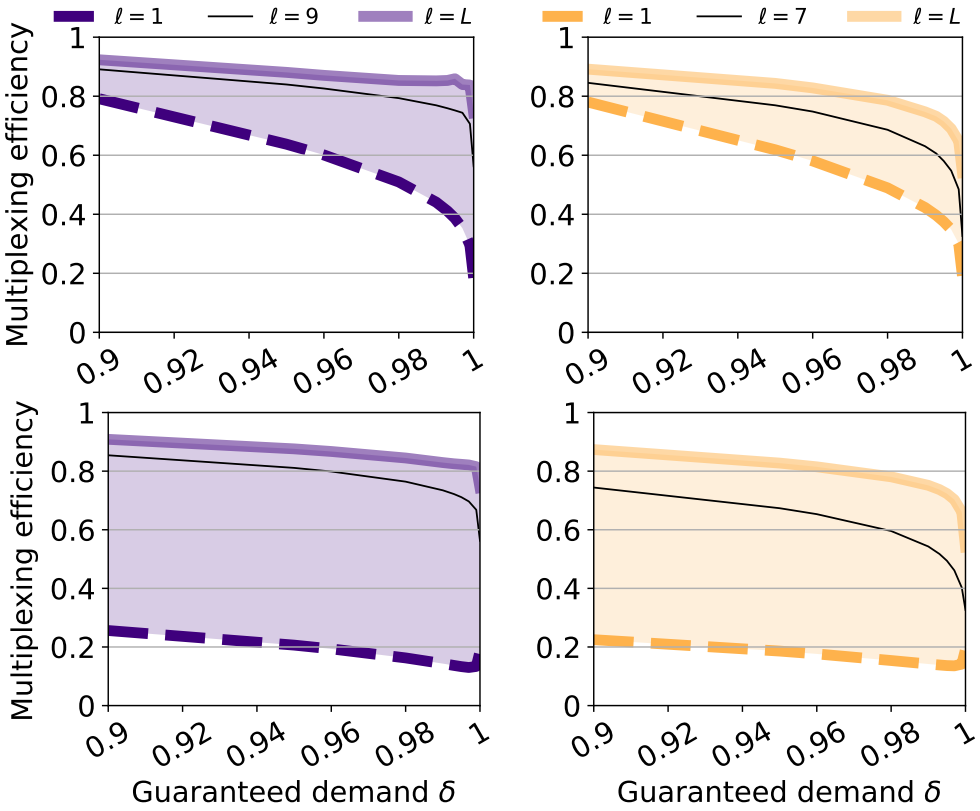


Figure 6.11. Efficiency of slice multiplexing versus slice specifications when $\pi = 0$.

Figure 6.11 offers a complete overview of sensible resource configuration schemes, in which we vary the overall QoS that each tenant is provided by the operator. It is important to remember that similar plots to this Figure will use the following format in this thesis. Thick dashed and solid lines represent the extreme network levels $\ell = 1$ and $\ell = L$, while thin solid lines are for an intermediate network level (*i.e.*, Mobile Edge Computing (MEC)), for the large metropolis (purple) and medium-sized city (gold).

Consistently with our system model, different QoS levels are reflected by diverse values of δ and π, hence we explore the impact of those two parameters on the multiplexing efficiency. The first configuration is depicted in the top pair of plots in Figure 6.11, which portray efficiency as a function of the guaranteed demand δ expressed as a time slot fraction, with no overbooking ($\pi = 0$). As one would expect, efficiency grows when not all the traffic demand for each slice has to be served. The increase is much more evident in the case of antenna-level resources ($\ell = 1$) than in the network core ($\ell = L$).

The good news is that a large fraction of the gain is achieved close to $\delta = 1$, *i.e.*, a slight reduction from a fully guaranteed demand may yield a large gain: in the best case, reducing δ from 1 to 0.99 raises efficiency from 0.35 to 0.6 (a 71% increase) when $\ell = 7$ in the medium-sized city scenario. The bad news is instead that efficiency values that are actually serviceable for the operator are only reached when significant amounts of traffic are not accommodated: figures above 0.8 (implying that implementing network slicing requires no more than 25% additional resources) are achieved in all configurations only when $\delta = 0.9$, and 10% of the demand is denied.

Trends are similar when the same slice specification parameters (varying δ, and $\pi = 0$) are defined as a traffic volume fraction, in the bottom pairs of plots in Figure 6.11. The major differences are at the antenna level ($\ell = 1$), where the multiplexing efficiency is substantially lower than in the case of δ and π expressed as time slot fractions. Indeed, imposing QoS constraints in terms of time slots or traffic volume leads to comparable efficiency when all time slots contribute a similar amount of traffic volume, and the demand is even over time. Centralized cases with high ℓ are closer to this situation.

However, we already noted in Section 6.6.1 that demands are much more irregular close to the radio access: here, most of the traffic volume is contributed by high activity peaks, and volume-based thresholds must still accommodate a significant portion of such peaks, instead of ignoring them completely as in the time slot-based case. Thus, volume-based service specifications at the antenna level force the operator to deploy a substantial amount of resources per slice even under more relaxed guaranteed demands.

Statistics are very different when including overbooking in the picture. Figure 6.12 illustrates the impact of the overbooking penalty (π), when the full demand is guaranteed, *i.e.*, $\delta = 1$. The plots refer again to pairs of scenarios, under slice specifications expressed in terms of time slots (top pair) and traffic volume (bottom pair). In almost all settings, the multiplexing efficiency quickly rises beyond 0.8 by just having 3% of the slice traffic not served in isolation. The only exception occurs for traffic volume-based guarantees at the antenna level: in this case, the efficiency gain with π is lower, yet the improvement with respect to the case where δ is varied (bottom pair in Figure 6.11) is dramatic.

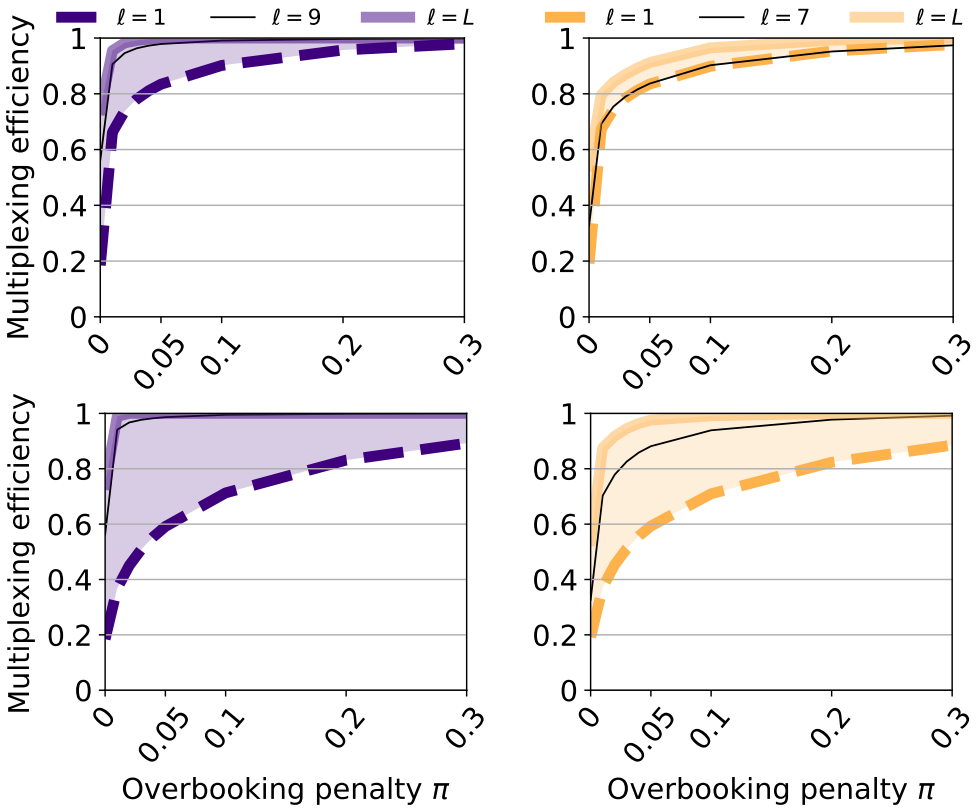


Figure 6.12. Efficiency of slice multiplexing versus slice specification when $\delta = 1$.

These results let us conclude that an overbooking that leads to serving a small portion of traffic peaks in a best-effort fashion is an interesting strategy for the operators, maintaining high standards ($\delta = 1$) with a reasonable increment of resources ($\leq 25\%$).

6.6.3 Slicing under dynamic resource orchestration

All previous results refer to cases where resources are statically allocated. We now investigate the multiplexing efficiency of network slicing when the operator can orchestrate network resources in an adaptive way, by re-allocating them to different slices over time.

As discussed in Section 6.3, this is equivalent to considering a resource reconfiguration interval τ that is shorter than the system observation time in our system model. Specifically, we assume that the operator can reconfigure the resources at each network level ℓ with a fixed periodicity τ which depends on the capabilities of the underlying virtualization technology. In our study, the operator allocates resources optimally to meet all slice specifications in each reconfiguration interval of duration τ. This is equivalent to assuming the availability of an oracle algorithm that, at the beginning of a reconfiguration interval, has perfect knowledge of the future demand for each service over the rest of the interval. Then, the operator can reserve for each slice the minimum amount of resources to abide by the requirements, as detailed in Section 6.3 and exemplified in Figure 6.2.

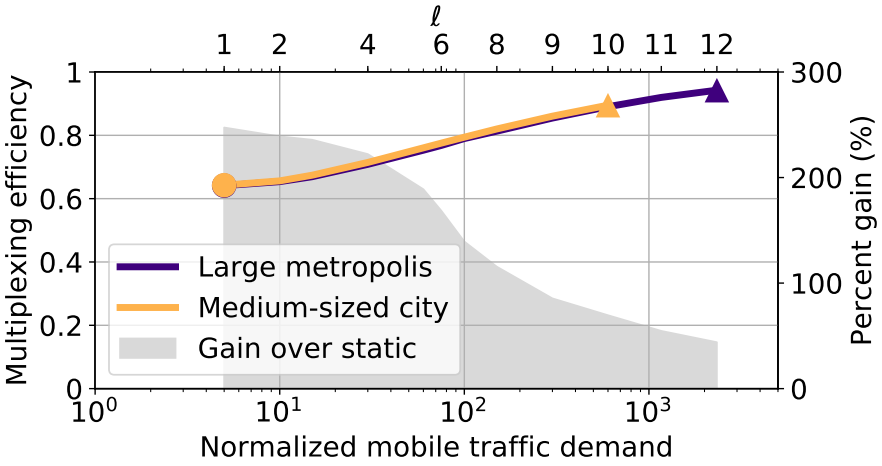


Figure 6.13. Efficiency of slice multiplexing (left y axis) and percent gain over static assignment (right y axis) versus the normalized mobile traffic served by one node (bottom x axis) at level ℓ (top x axis) in the two reference urban scenarios.

Our baseline result, in Figure 6.13, refers to $\tau = 30$ minutes, which can be regarded as a fairly high resource reconfiguration frequency for several scenarios. For instance, VNF management in the network core cloud has typically larger time scales of hours or even days [90]. At radio access, instead, faster dynamic reassignments are technically possible; however, forecasting the demand over short time scales of minutes is challenging and easily leads to slice specification violations, hence reconfiguration intervals in the order of hours are more credible [41].

In these settings, dynamic allocation mechanisms and a perfect prediction of the demand over the future 30 minutes can substantially improve the efficiency of slice multiplexing. Indeed, when comparing the curves in Figure 6.13 (under slice specification $z = (1, 0)$) with their equivalent in Figure 6.9, the gain is evident. We explicitly portray the benefit as the grey region in Figure 6.13: it ranges between 90% ($\ell = L$) and 250% ($\ell = 1$). The cause of such a significant advantage roots in that different mobile services allocated to separate slices tend to peak at different times of the day, as discussed in details in recent analyses of mobile service dynamics [1]. The temporal diversity of peaks across slices lets a perfect orchestrator reuse the same resources to cover time-disjoint high-activity periods in multiple slices, hence increasing the system efficiency.

Despite the much higher gain at the antenna level, there is still a large gap between the efficiency at the radio access and in the network core. An order-of-minute dynamic orchestration of resources allows for near-perfect slice multiplexing at a datacenter that fully centralizes the traffic in our large metropolis scenario. In contrast, efficiency is bounded at around 0.6 for levels close to $\ell = 1$, *i.e.*, at individual antenna sectors or at nodes serving small groups of a few antennas each. This implies that the operator still has to nearly double the capacity to isolate slices at network levels close to the radio access.

A more comprehensive picture is provided by Figure 6.14, which encompasses a wide set of reconfiguration intervals τ, from the 30-minutes case we just analyzed in detail up to 3 months, *i.e.*, the entire timespan of the dataset, which maps to the static resource configuration case considered in Section 6.6.1. As one could expect, the multiplexing efficiency of slices is decreased as τ grows, since the system becomes

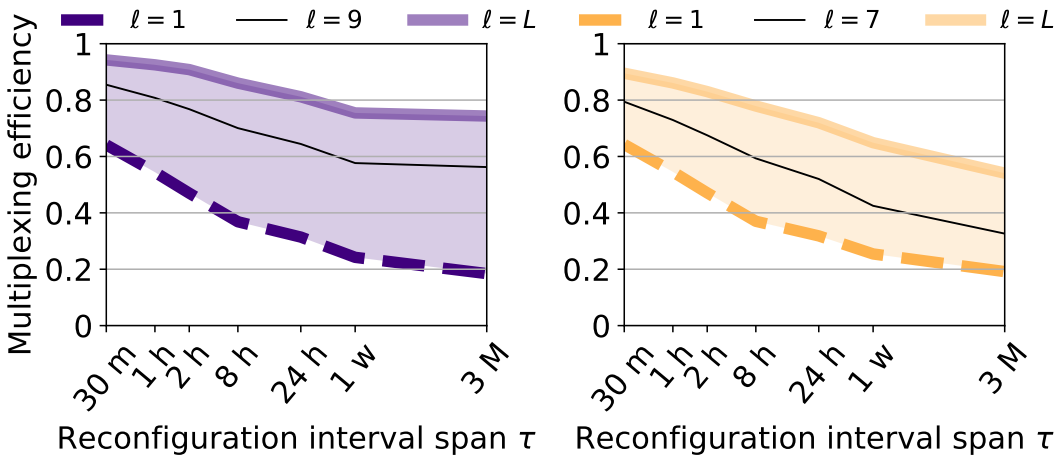


Figure 6.14. Efficiency of slice multiplexing versus the resource reconfiguration periodicity τ. Left: large metropolis. Right: medium-sized city.

less flexible. Interestingly, the loss of efficiency is steeper at lower values of τ: reducing the frequency of reallocation from once every 30 minutes to once every day yields an efficiency loss comparable to that caused by increasing τ from one day to 3 months. This is consistent with the typical duration of human activities, in the order of tens of minutes, which reflects on similar timescales of mobile service demand fluctuations [1]. Therefore, predicting traffic and allocating resources at longer periodicity rapidly reduces the system efficiency: either the operator is able to deploy virtualization technologies that enable such a reconfiguration frequency, or it is probably not worth considering dynamic resource allocation at all.

6.6.4 Varying number of slices

Up to now, we assumed that the slicing strategy adopted by the operator involved assigning one slice to each of the 16 services that generate the most traffic. In fact, the mapping of services into specific network slice instances is a business-driven choice that is based on several factors, such as the requirements of the services in terms of isolation, the specific policies implemented by the operator [91], or the practice of the tenants, which may decide to group multiple services into a same slice for economic reasons. The number of slices and the demands associated to each will have an impact on the overall multiplexing efficiency, which we investigate next.

We first analyze a business-driven scenario where network slices are dedicated to sets of services of a same category, *i.e.*, streaming, social media, web, cloud, gaming, messaging and miscellanea, respectively. Here, we set $\delta = 1$ and $\pi = 0$ for all slices. In this scenario, we study the impact of the system reconfiguration dynamics, as displayed in Figure 6.15. Trends are similar to those observed for a per-service slicing in Figure 6.14. Despite a higher efficiency in general, the fractional gain brought by increasingly faster resource orchestration is comparable under the two different slicing policies.

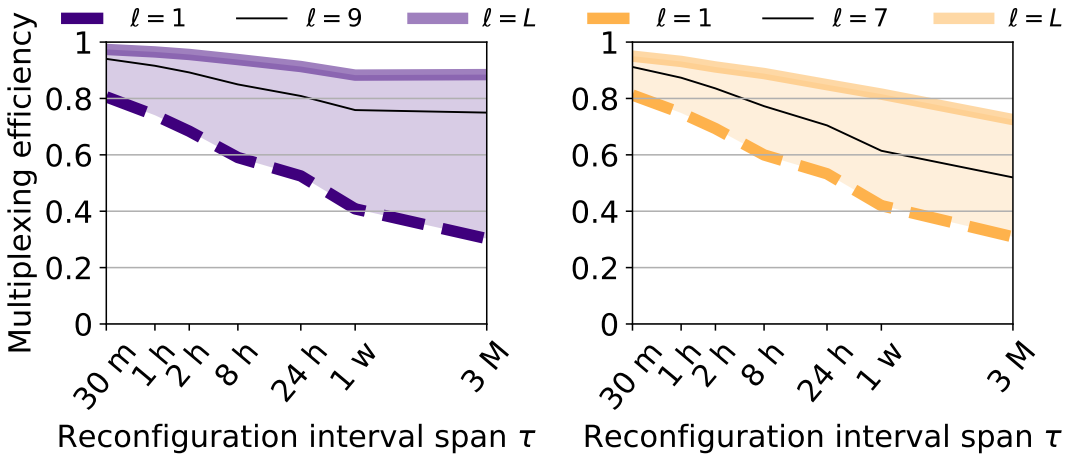


Figure 6.15. Efficiency of slice multiplexing with per-category slicing. The plot semantics are the same as in Figure 6.14.

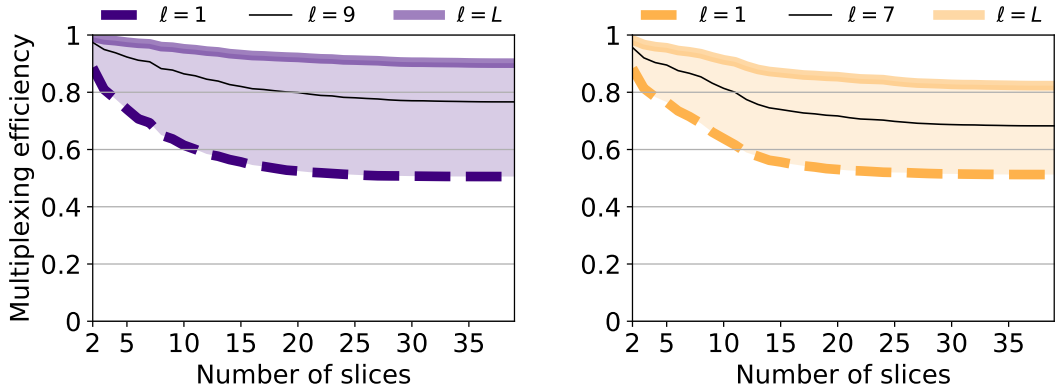


Figure 6.16. Efficiency of slice multiplexing as a function of the number of slices $k+1$ (on the x axis), when the k services with the highest traffic load have a dedicated slice and the remaining services are aggregated into a common slice.

We then explore different slicing strategies according to a hierarchical scheme where the k services that generate the highest traffic loads acquire a dedicated slice each. The demands for all remaining services are instead aggregated into a common, non-customized, slice. Figure 6.16 shows the resulting multiplexing efficiency as a function of the total number $k+1$ of slices in the network, when the reconfiguration period τ is set to 1 hour, $\delta = 1$ and $\pi = 0$ for all slices. Increasing the number k of isolated mobile services entails a reduction of efficiency: this is expected, since a larger k moves traffic from the common slice, within which multiplexing is perfect, to dedicated slices that require isolated resources. Interestingly, however, the loss of efficiency is accumulated in the first half of the plots, *i.e.*, considering a number of slices larger than 16 does not affect efficiency anymore. Therefore, most of the resource utilization cost for the operator comes from the very few mobile services that generate the largest demands, and multiplexing efficiency is only increased when such services are treated as best-effort traffic. Incidentally, these results also motivate our choice of focusing on 16 slices in previous experiments: this setting maps to a lower bound on performance in terms of efficiency with our dataset.

A second sensible slice configuration assumes that the providers of the services that generate the highest traffic load acquire a dedicated slice tailored to their service, while the remaining services are aggregated into a common, non-customized, slice. In Figure 6.16, we analyze the multiplexing efficiency resulting from this configuration as a function of the total number of slices in the network (including the dedicated slices and the common one) when the reconfiguration period τ is of 1 hour and $f = 1$ for all slices. Results show that the trend becomes almost flat after 15 slices, which implies that efficiency is only improved when the services with the largest demands are brought into the common slice.

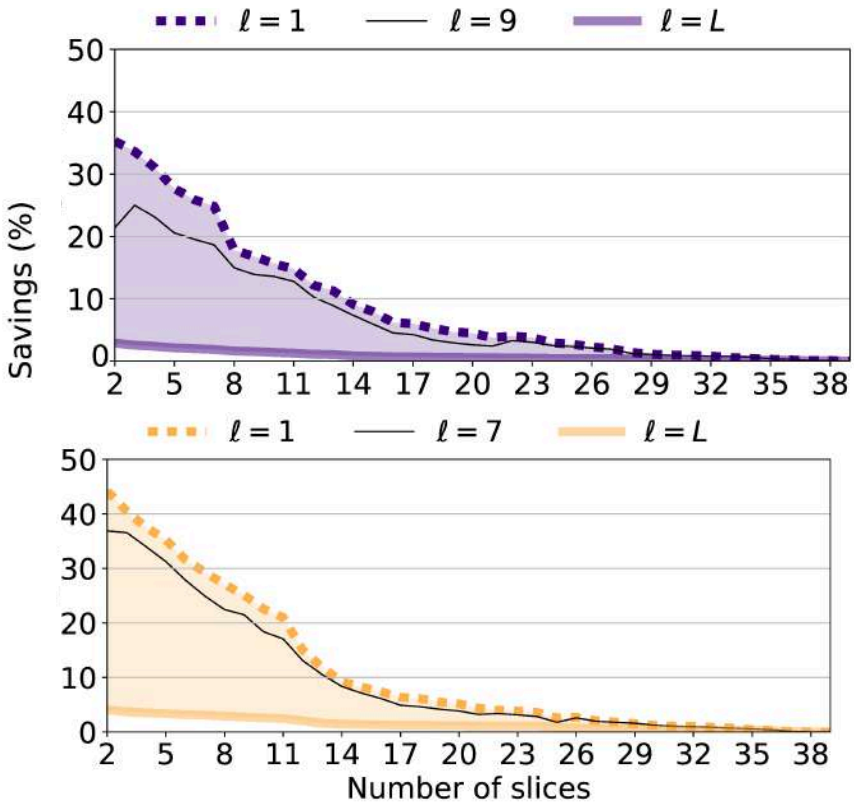


Figure 6.17. Savings obtained by relaxing the service guarantees of the common slice, corresponding to the difference between the resources required when $f = 1$ for the common slice, and those required when $f = 0.9$ for that slice.

In the above slice configuration, it may be reasonable to expect that those tenants acquiring dedicated slices are provided a stricter guarantees than the ones in the common slice. In order to evaluate the benefits resulting from such a strategy, Figure 6.17 illustrates the resource savings resulting from providing the common slice with a guaranteed time fraction $f = 0.9$, computed as the relative percentage of resources spared with respect to those required in the configuration where all slices have $f = 1$. Results show that savings remain very low in the network core (when $\ell \sim L$), but can be significant for resources located close to the radio access (when $\ell \sim 1$). In the latter case, savings are important (up to 20-40%) when the top-10 services are included in the non-customized, low-QoS common slice. Indeed, as these account for 65% of the overall traffic (see Figure 6.5), they have a much higher incidence on the system performance.

6.6.5 Case studies

In our evaluation, we have studied each system parameter in isolation. In this section, we investigate the multiplexing efficiency under network slicing in a number of specific case studies. This analysis lets us detail particular settings of practical interests, and complements the previous results. Each case study focuses on a specific service category (*e.g.*, video streaming), where we assume that different applications (*e.g.*, YouTube, iTunes, DailyMotion, NetFlix, etc.) are allocated to isolated slices. The detailed configurations and the associated efficiency results are provided in Table 6.2, for both the large metropolis (LM) and medium-sized city (MC) scenarios.

Category	Slices	Network level	τ	Slice specification			Efficiency	
				Guarantee	δ	π	LM	MC
Streaming	**8**	**Antenna ($\ell = 1$)**	**1h**	**Volume**	**1**	**0**	**0.35**	**0.35**
Streaming	8	MEC ($\ell = 9, 7$)	4h	Volume	1	0	0.74	0.59
Streaming	8	Core ($\ell = L$)	10h	Volume	1	0	0.84	0.73
Web	**6**	**Antenna ($\ell = 1$)**	**1h**	**Volume**	**1**	**0.02**	**0.71**	**0.71**
Web	6	MEC ($\ell = 9, 7$)	4h	Volume	1	0.02	0.97	0.85
Web	6	Core ($\ell = L$)	10h	Volume	1	0.02	1	0.96
Social media	**4**	**Core ($\ell = L$)**	**10h**	**Time slot**	**0.99**	**0.05**	**0.90**	**0.96**
Social media	4	Antenna ($\ell = 1$)	1h	Time slot	0.99	0.05	0.81	0.81
Social media	4	MEC ($\ell = 9, 7$)	4h	Time slot	0.99	0.05	0.92	0.89
Gaming	**6**	**MEC ($\ell = 9, 7$)**	**4h**	**Volume**	**1**	**0**	**0.57**	**0.59**
Gaming	6	Antenna ($\ell = 1$)	1h	Volume	1	0	0.58	0.59
Gaming	6	Core ($\ell = L$)	10h	Volume	1	0	0.57	0.65
Messaging	**5**	**MEC ($\ell = 9, 7$)**	**4h**	**Volume**	**0.99**	**0.03**	**0.68**	**0.80**
Messaging	5	Antenna ($\ell = 1$)	1h	Volume	0.99	0.03	0.5	0.45
Messaging	5	Core ($\ell = L$)	10h	Volume	0.99	0.03	0.91	0.89

Table 6.2
Case studies. Each row maps to one configuration.

Our analysis below addresses one network level in each case study, highlighted in bold in Table 6.2. For the sake of completeness, the Table also includes additional levels for each scenario, which allow appreciating, for each case study, the efficiency of end-to-end slicing across the network architecture.

Case study #1 – High QoS at the access network. The first case study focuses on slicing at the antenna level, and on capacity-demanding services such as video streaming and web access. These are challenging settings for the operator, who must provide high-quality support for a large volume of bursty traffic; a quite fast reconfiguration ($\tau = 1$ h) is thus a reasonable relief. The efficiency is nonetheless low if hard-QoS requirements ($\delta = 1$) are to be met, *e.g.*, for video streaming slices: the operator shall commit up to threefold the resources needed in a non-sliced scenario – a high cost considering that radio access resources such as spectrum or RAN processing capacity can be very expensive. In less strict slices like those dedicated to web access, paying minimal overbooking penalties ($\pi = 0.02$) is an appealing option, as it may reduce costs considerably by raising efficiency to 0.71.

Case study #2 – Large traffic flows in the core. This case study shifts the focus to datacenters in the network core, and targets social media services that generate high demands but are less latency-dependent. The large traffic volumes observed at this level allow achieving high efficiency (above 0.90) under loose QoS ($\delta = 0.99, \pi = 0.05$), and with limited reconfiguration possibilities ($\tau = 10$ h). These results further prove the benefits of centralization for the effective implementation of network slicing.

Case study #3 – Computing at the edge. Gaming services with strong QoS requirements are likely candidates to be among the first services to be delivered over edge deployments [92], hence they represent a sensible target for MEC-level slicing.

We consider 6 popular mobile games, to which we allocate dedicated slices with firm specifications ($\delta = 1$, $\pi = 0$). Although we allow for quite fast reconfiguration (τ = 4h), the price that the operator has to pay is high in both urban scenarios: the required resources are almost doubled with respect to a non-sliced network. Similar considerations hold for messagging services, although in this case QoS requirements can be moderated to $\delta = 0.99$ $\pi = 0.03$, with a 20-30% efficiency gain with respect to the gaming case.

6.6.6 Equipment deployment efficiency

To conclude our analysis, we look at the problem of resource multiplexing efficiency in a sliced network from a rather different perspective. The expressions (6) and (9) derived in Section 6.4 assume that the relevant metric for the operator is the amount of resources utilized to accommodate the demand for mobile services aggregated over time. Therefore, the analysis carried out in Sections 6.6.1–6.6.5 is appropriate to evaluate OPerating EXpense (OPEX), which increase when the available resources are used more intensively, and can be applied, *e.g.*, to electric power consumption, management overheads, or deterioration of assets with use.

However, another interesting viewpoint is the efficiency in terms of equipment to be deployed to meet the instantaneous demand. This relates to the CAPital EXpenditure (CAPEX) incurred by the mobile network operator, typically hardware and infrastructure. In this case, the cost expressions are slightly different, and capture the fact that the equipment must be dimensioned so as to match the peak demand. Formally, let $\hat{r}^{\mathbb{z}}_{c,s}(n)$ be the resources needed to satisfy specifications $\mathbb{z}$ for slice $s \in \mathcal{S}$ at node $c \in \mathcal{C}_\ell$ during reconfiguration interval $n \in \mathcal{T}$, computed as indicated in Section 6.3. Then, the equipment resources needed to accommodate the traffic activity peak in slice s at network level ℓ are computed as

$$\mathbb{D}^{\star\mathbb{z}}_{\ell,\tau} = \sum_{s\in\mathcal{S}} \sum_{c\in\mathcal{C}_\ell} \max_{n\in\mathcal{T}} \left(\hat{r}^{\mathbb{z}}_{c,s}(n)\right). \tag{19}$$

Similarly, the equivalent resources needed under perfect sharing in the same settings are

$$\mathbb{P}^{\star\delta}_{\ell,\tau} = \sum_{c\in\mathcal{C}_\ell} \max_{n\in\mathcal{T}} \left(\hat{r}^{\delta}_{c}(n)\right), \tag{20}$$

where $\hat{r}^{\mathbb{z}}_{c}(n)$ is the amount of resources needed to accommodate the total demand aggregated over all slices in $\mathcal{S}$ at node c and reconfiguration interval n, under requirements $\mathbb{z}$. The multiplexing efficiency for deployed equipment is then

$$\mathbb{E}^{\star\mathbb{z}}_{\ell,\tau} = \mathbb{P}^{\star\delta}_{\ell,\tau} / \mathbb{R}^{\star\mathbb{z}}_{\ell,\tau}. \tag{21}$$

The equipment deployment efficiency in (21) is shown in Figure 6.18. The figure summarizes results in our reference urban scenarios, under a wide range of reconfiguration time interval durations τ, and across all network architectural levels ℓ. We highlight the following aspects.

(i) In absence of mechanisms that allow for dynamic reconfiguration, the efficiency is very much comparable to that observed in the previous analysis, as shown by the values for τ = 3 months in Figures 6.14 and 6.18. This is a clear indication that deploying hardware and infrastructure to provide resource isolation across slices risks to have an unbearable cost for operators if no dynamic resource reallocation is possible.

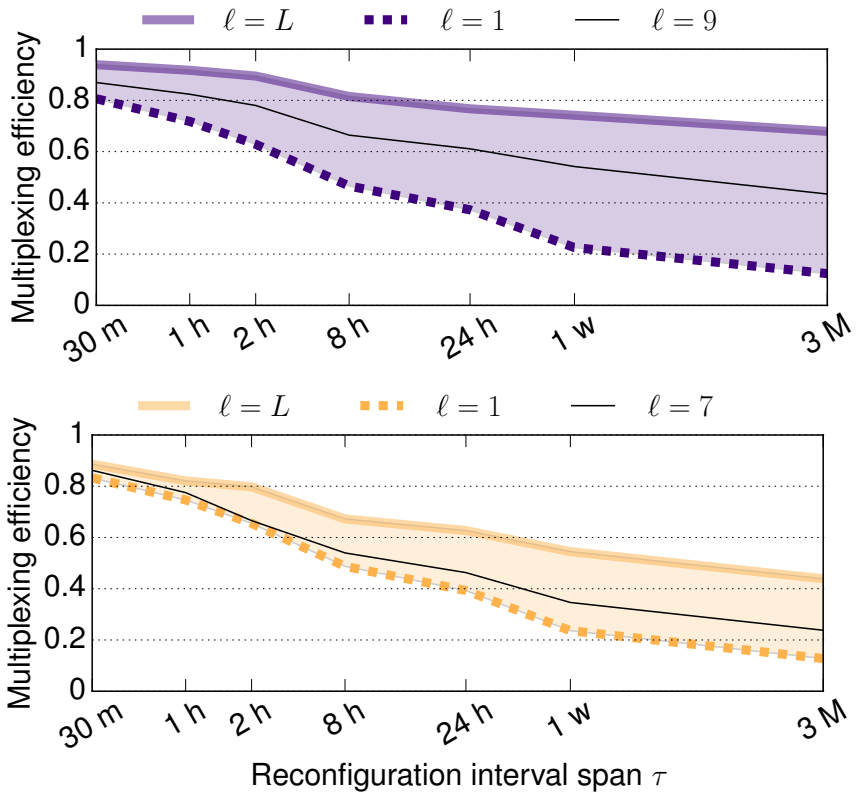


Figure 6.18. Efficiency of slice multiplexing for an equipment deployment perspective versus the resource reconfiguration periodicity τ.

(ii) Flexibility in the orchestration of resources pays off also in terms of equipment deployment efficiency, which can be increased up to 0.8–0.95 when fast reconfiguration over 30-minute intervals is possible. These values correspond to an additional 5%–25% cost in terms of network infrastructure over the perfect sharing benchmark.

(iii) The main difference between efficiency of resource usage, given by (10), and equipment deployment, given by (21), is observed at architectural levels closer to radio access. When ℓ is close to 1, a dynamic reconfiguration of resources allows improving deployed infrastructure efficiency much faster than resource usage efficiency. In other words, resource isolation across slices has a sensibly lower impact on equipment installation costs than on operating expenses. For instance, at the antenna level ($\ell = 1$), efficiency is 0.6 in Figure 6.14 and 0.8 in Figure 6.18, implying that the extra cost over perfect sharing is high for resource utilization (over 60%) and much lower for equipment deployment (below 25%).

(iv) In contrast to the above, in the network core (*i.e.*, for ℓ that tends to L) trends are similar in Figure 6.14 and Figure 6.18.

Overall, our results stress how multiplexing efficiency of slice resources is largely consistent across the different perspectives entailed by the expressions (10) and (21). That is, the OPEX and CAPEX incurred by the operators to support network slicing have comparable trends with respect to the different system parameters, with the notable exception of lower deployment costs for a radio-access infrastructure supporting high reconfigurability.

6.7 Takeaways

Our data-driven analysis unveils how real-world service usage patterns may affect the deployment of a key paradigm for future-generation mobile networks such as network slicing, and the impact it has on resource management. Specifically, we retain the following main takeaway messages:

Multi-service requires more resources. Building a network that is capable of providing different services (possibly associated to several tenants) will necessarily reduce efficiency in resource utilization. We quantify this loss in almost one order of magnitude if considering distributed resources (such as spectrum), yet the efficiency loss stays as high as 20% even in large datacenters in the core network. These figures translate into high costs for the infrastructure provider, who must compensate for them by aggressively monetizing on the new business models enabled by a multi-service scenario (*e.g.*, Network Slice as a Service, Infrastructure as a Service).

Traffic direction is a factor. Uplink and downlink traffic exhibit similar efficiency trends across network levels, but uplink exacts a much higher efficiency degradation to meet equivalent QoS requirements. Although uploads account for a small fraction of the overall load, the lower efficiency of uplink may entail additional challenges for the operators. Indeed, uplink QoS requirements are key to specific services such as mobile gaming, and it is likely that multiple instances of such services belonging to different tenants have to be served in a resource-isolated fashion in parallel.

Loose service level agreements may not help. Although slice specifications may be moderated, the overall efficiency grows only when guarantees on the serviced demand are very much lowered, up to a point that they may not be suitable for certain services (needing, *e.g.*, "5 nines reliability", or strict bandwidth requirements over very short time slots).

Overbooking is a key strategy. While downgrading the requirements in terms of served fraction of traffic only helps when brought to extreme levels, flexibly serving small portions of the individual slice demands via a non-customized common slice provides high benefits. Therefore, overbooking solutions that only marginally underserve slices may yield substantial economic gains for the operators, as they allow trading off substantial resource deployment costs with negligible penalty fees due to slight SLA violations. This corroborates the importance of recent approaches for practical end-to-end resource overbooking in sliced 5G networks [84].

Guaranteeing traffic volumes at the antenna is costly. If operators define SLAs in terms of assured traffic volumes, they shall note that meeting the QoS requirements will need substantial additional resources at the radio access, even if guarantees are loose and overbooking is in place. SLAs defined in terms of guaranteed time slots allow much more flexibility in balancing efficiency and QoS for each network slice.

Dynamic resource assignment must be rapid. The design of dynamic resource allocation algorithms is crucial to increase the efficiency of future sliced networks. However, substantial gains will only be attained if the virtualization technologies enable a fast enough re-orchestration of network resources. While current Management and Orchestration (MANO) frameworks provide such capabilities, intelligent algorithms able to forecast mobile service demands and anticipate resource reconfiguration are also required, Artificial Intelligence (AI) and Machine Learning (ML) are promising techniques to accomplish this [93], and are also being brought into the network management landscape by standards [94].

Aggregating services is beneficial. Aggregating similar services into a same slice increases the system efficiency significantly, yet it comes at the price of losing the ability to customize treatment to each service. This implies that operators may face a business trade-off between providing dedicated support to highly remunerative, popular services, and incurring high management costs to implement the associated slices.

Deployment is slightly more efficient than operation. We analyzed the sharing efficiency from both a continuous resource usage and an infrastructure deployment perspective. While they have similar trends in the network core, the efficiency at the radio access is higher for installed hardware in presence of high-frequency resource reallocation.

Urban topography has limited impact. The fact that all of our results are very consistent in two urban areas with a quite different nature lets us provide general insights that hold beyond one particular scenario. More precisely, as usage demands are eventually driven by human factors, we expect that our considerations might apply to other metropolitan regions in (and possibly beyond) Europe.

Efficiency under uncertain load demands. Our analysis concerns resource management efficiency under known loads, as slices are allocated the exact resources needed to meet the corresponding service demands. This lets us investigate the impact of the limited reconfigurability of resources, which forces the operator to provision a constant amount of resources during the following reconfiguration period. In a real system, however, the network slices demands are not known *a priori*, and resources have to be allocated based on a forecast of the expected demand during the next re-orchestration interval. This introduces a second source of inefficiency, *i.e.*, the inaccuracy of traffic predictions, which imposes some overprovisioning in the allocated capacity to combat the uncertainty associated with the future load information. This second aspect has been recently analyzed by the authors in [95], where an approach is developed that forecasts the capacity needed to accommodate the traffic of a slice. Figures about the expected global performance of a practical system can then be obtained by summing the effects of both sources of inaccuracy. For instance, if the resource reconfiguration periodicity imposes allocating 100% extra resources (which is a typical case according to the results in the previous sections), and capacity predictors entail 10% overprovisioning (a likely number according to [95]), then the overall additional resources required will amount to 110%. This extra capacity can then be served with a mixture of guaranteed demand and overbooking, as discussed in Section 6.1.

7

Conclusions and future work

Contents

This thesis has analyzed several factors involving the mobile service usage and network slicing models in diverse scenarios. Particularly, the clusterization of mobile services in distinct domains (*i.e.*, the temporal, spatial, spectral and wavelet approaches), and the network trade-offs of network slicing according to different parameters (such as the guaranteed time fraction, the reconfiguration interval span or the number of slices, level of aggregation, or the percentage of overbooked resources) in two urban scenarios have been addressed.

This chapter synthetizes the contributions in this thesis and presents the main conclusions as follows. The first section presents the conclusions of the research. The second section details the possible future work based on the findings of this thesis.

7.1 Conclusions

As the volume of data, digital transformation, and the pace of technological change accelerate, the ability of organizations and professionals to keep up and capitalize on the opportunity is becoming more challenging. In particular, traditional software-based approaches cannot deal with the new heterogeneity of service's demands in the 5G paradigm. It is in this framework where SDN and NFV technologies appear, as part of the new set of equipment and techniques needed. Among them, network slicing seems to be the most promising tool for the allocation of needed resources when customization of services, KPIs and QoS guarantees are essential.

Given this new reality, operators and tenants will work in a multi-domain network context, where adaptability and programmability are paramount, as well as data isolation and management automation. Our data-driven study contemplates an extensive dataset with a high-granularity, discussing two main areas that will help decision makers to base their investments in real-world information and address issues before they become problems.

On the one hand, we identify the following classes of components in the time series of the demands of multiple popular mobile services: (*i*) *non-composable*, *i.e.*, recurrent patterns that are found in all time series, hence are inherently impossible to compose (their sum is just a scaling of the pattern observed in each time series); (*ii*) *composable*, *i.e.*, recurrent patterns that are specific to each time series (which are the portion we can actually sum with some hope to obtain near-constant load); and (*iii*) *noise*, non-recurrent patterns that are excluded from our analysis, small in magnitude. Then, we correlate and apply clustering algorithms to the relevant features in three dimensions (*i.e.*, time, space and frequency) to classify the service's behaviour and be able to provide recommendations on how to allocate network resources for distinct clusters.

A first finding from this research is that no two services exhibit similar time patterns in their nationwide aggregate traffic: although expected for different service categories, this is less obvious for akin services, e.g., diverse applications that all provide video streaming. A second key insight is that mobile services have very comparable geographical distributions of both total and the per-user traffic demands. That is, different services have different temporal patterns (*i.e.*, they are consumed at different times), but their geographical patterns (*i.e.*, locations where they are consumed) are very similar. Our third takeaway message is that spatial distributions of per-subscriber service usage are in fact driven by land use, *i.e.*, the urbanization level plays a major role in influencing how much mobile services users consume. Nonetheless, it has a much lower impact on when they do so, as the average subscribers in urban, semi-urban and rural regions all follow similar service access patterns; a notable exception is represented by users on high-speed trains, who show unique time dynamics. Along these lines, we could identify common periodic behaviors in the real-world traffic generated by a large set of applications by leveraging spectral methods.

On the other hand, we carry a profound analysis on quantifying resource management efficiency and cost-effectiveness of the system showing the trade-off between *(i)* assigning dedicated resources for service customization purposes, and *(ii)* resource sharing practices among services. Particularly, our results provide insights on the achievable efficiency of network slicing architectures, their dimensioning, and their interplay with resource management algorithms at different locations and reconfiguration timescales.

First, we prove that network providers will face challenges in scenarios where the reconfiguration of their networks should be fast (*i.e.*, at timescales shorter than a few hours), drastically reducing the time to act, for an increasing number of mobile services. Second, we observe that traffic direction is a factor, and it should be taken into account when deciding where to increase or decrease the number of slices (*e.g.*, aggregating services only has an effect when the top-10 services are concerned in the downlink direction). We remark that a multi-service scenario in general requires more resources, but the monetary investment and the gains related to network slicing usage depend on the network level (*e.g.*, antenna or cloud level).

Moreover, we scrutinize that QoS specifications must be reduced vastly to achieve high efficiency. Even loosing SLAs may not help, as efficiency could be achieved when requirements may not be suitable for certain services. However, overbooking is a key strategy to increase operators revenues with slight penalty fees. Last but not least, our results are consistent in the audited urban scenarios, meaning that the impact of urban topologies is limited.

While a thorough analysis of the overall efficiency resulting from considering both effects is left as future work, it is worth mentioning that, according to the results presented in this thesis and in [95], it is expected that the overall efficiency will be dominated by the resource allocation dynamics analyzed in this thesis.

To conclude, ours does not pretend to be a fully comprehensive analysis, rather one that lays the foundations to a better understanding of the new trade-offs introduced by network slicing in terms of resource management efficiency. The empirical bounds we derived represent a starting point for deeper investigations of an unexplored subject with strong implications for the future generations of mobile networks.

7.2 Future work

This work provides an original perspective on the temporal analysis of mobile applications. By leveraging distinct methods, we could identify common behaviors in the real-world traffic generated by a large set of services, which were not detected by previous studies. Our results pave the road for further investigations, aimed at explaining the root causes for these temporal similarities, at assessing their generality at different spatial and geographical scales, and at exploiting them for applications in network planning and resource management. For instance, new and improved data plans, widespread broadband availability, and services anticipate and meet the demands, will boost new lifestyles where connectivity and mobility are paramount.

Besides, the proposed methodology to generate and work with synthetic data solves privacy derived issues stated in the GDPR. The combination of this methodology, the hybrid time-frequency approach, and forecasting algorithms on data traffic consumption becomes an opportunity to help professionals close the gap and harness the full potential of data, creating new tools to improve their network infrastructure and outcomes. Consequently, the risk of over-provisioning would be reduced.

However, our work is meaningful for domains beyond networking, as it will help to establish a symbiosis between the Telecommunications and Transportation industries to improve the mobile coverage and the transport connection at under served regions. Hence, society will have additional tools to fight social segregation issues and avoid self-contained residential areas. In addition, as mentioned before, when new mobility and lifestyle trends appear, it will affect other fields of study such Sociology, Education or Medicine. For example, the digital transformation will modify how we work and study, or it will allow us to receive telesurgery treatments and telemedicine diagnosis.

On a side note, as the dataset is from 2016, it would be interesting to update the information up to the present day and study the main changes found between [1] and [4], as well as to perform historical evaluations in different countries that consume other mobile services (*e.g.*, Line messaging service). Also, we can integrate current algorithms in systems and work on the algorithms that leverage our analysis, while we keep track on how the parameters defined in our models evolve with respect to them.

Of course, this is out of the scope of this thesis, but it would be of significant value as it would allow to study both the composability of resources, and the flexible SLA adherence for future 5G networks. For example, we could quantify the network efficiency of resources when clustering together services that are synchronized in distinct regions, as they tend to consume more resources over the same period, while this effect could not be exhibited by the non-clustered ones.

References

[1] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in *Proceedings of the 13th International Conference on Emerging Networking EXperiments and Technologies (ACM CoNEXT 2017)*, Incheon/Seoul, Republic of South Korea, Dec. 2017, p. 180–186. [Online]. Available: https://doi.org/10.1145/3143361.3143369

[2] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and Z. Smoreda, "Identifying Common Periodicities in Mobile Service Demands with Spectral Analysis," in *Proceedings of the 18th Mediterranean Communication and Computer Networking Conference (IEEE MedComNet 2020)*, Arona, Italy, Jun. 2020, pp. 1–8. [Online]. Available: https://doi.org/10.1109/MedComNet49392.2020.9191477

[3] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Pérez, "Resource sharing efficiency in network slicing," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 909–923, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8737701

[4] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How Should I Slice My Network? A Multi-Service Empirical Evaluation of Resource Sharing Efficiency," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2018)*, New Delhi, India, 2018, p. 191–206. [Online]. Available: https://dl.acm.org/doi/10.1145/3241539.3241567

[5] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the Decomposition of Cell Phone Activity Patterns and Their Connection with Urban Ecology," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing (ACM MobiHoc 2015)*. Hangzhou, China: Association for Computing Machinery, Jun. 2015, p. 317–326.

[6] E. Peltonen, E. Lagerspetz, J. Hamberg, A. Mehrotra, M. Musolesi, P. Nurmi, and S. Tarkoma, "The Hidden Image of Mobile Apps: Geographic, Demographic, and Cultural Factors in Mobile Usage," in *Proceedings of the 20th ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (ACM MobileHCI 2018)*, Barcelona, Spain, 2018.

[7] N. Oliver, B. Lepri, H. Sterly, R. Lambiotte, S. Deletaille, M. De Nadai, E. Letouzé, A. A. Salah, R. Benjamins, C. Cattuto, V. Colizza, N. de Cordes, S. P. Fraiberger, T. Koebe, S. Lehmann, J. Murillo, A. Pentland, P. N. Pham, F. Pivetta, J. Saramäki, S. V. Scarpino, M. Tizzoni, S. Verhulst, and P. Vinck, "Mobile phone data for informing public health actions across the covid-19 pandemic life cycle," *Science Advances*, vol. 6, no. 23, 2020. [Online]. Available: https://advances.sciencemag.org/content/6/23/eabc0764

[8] J. Alvarez-Lozano, V. Osmani, O. Mayora, M. Frost, J. Bardram, M. Faurholt-Jepsen, and L. V. Kessing, "Tell Me Your Apps and I Will Tell You Your Mood: Correlation of Apps Usage with Bipolar Disorder State," in *Proceedings of the 7th*

ACM International Conference on PErvasive Technologies Related to Assistive Environments (ACM PETRA 2014), Rhodes, Greece, 2014.

[9] L. Ravindranath, S. Agarwal, J. Padhye, and C. Riederer, "Procrastinator: Pacing Mobile Apps' Usage of the Network," in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (ACM MobiSys 2014)*, Bretton Woods, NH, 2014, p. 232–244.

[10] M. De Nadai, J. Staiano, R. Larcher, D. Quercia, N. Sebe, and B. Lepri, "The death and life of great italian cities : A mobile phone data perspective," in *The World Wide Web Conference (ACM WWW 2016)*, Montréal, QC, Canada, Apr. 2016.

[11] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, and Z. Smoreda, "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, Oct 2017.

[12] D. Elias, F. Nadler, J. Stehno, J. Krösche, and M. Lindorfer, "Somobil – improving public transport planning through mobile phone data analysis," *Transportation Research Procedia*, vol. 14, pp. 4478 – 4485, 2016, transport Research Arena TRA2016.

[13] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[14] W. Roh, J. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

[15] D. Kim, J. Gluck, M. Hall, and Y. Agarwal, "Real World Longitudinal iOS App Usage Study at Scale," 2019.

[16] L. Qi, Y. Qiao, F. B. Abdesslem, Z. Ma, and J. Yang, "Oscillation Resolution for Massive Cell Phone Traffic Data," in *Proceedings of the First Workshop on Mobile Data (ACM MobiData 2016)*, Singapore, Singapore, 2016, p. 25–30.

[17] Q. Xu, A. Gerber, Z. M. Mao, and J. Pang, "AccuLoc: Practical Localization of Performance Measurements in 3G Networks," in *Proceedings of the 9th ACM International Conference on Mobile Systems, Applications, and Services (ACM MobiSys 2011)*, Bethesda, Maryland, USA, 2011, p. 183–196.

[18] 5th Generation Americas' (5G Americas'), "4G Americas' Summary of Global 5G Initiatives," 5G Americas' Working Group White Paper, Jun. 2014.

[19] C. Wang, F. Haider, X. Gao, X. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[20] P. Rost *et al.*, "Network slicing to enable scalability and flexibility in 5g mobile networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, 2017. [Online]. Available: http://arxiv.org/abs/1704.02129

[21] S. Sezer, S. Scott-Hayward, P. K. Chouhan, B. Fraser, D. Lake, J. Finnegan, N. Viljoen, M. Miller, and N. Rao, "Are we ready for SDN? Implementation challenges for software-defined networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 36–43, Jul. 2013.

[22] Y. Zhang and A. undefinedrvidsson, "Understanding the Characteristics of Cellular Data Traffic," in *Proceedings of the 2012 ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design (ACM CellNet 2012)*, Helsinki, Finland, 2012, p. 13–18.

[23] H. Wang, Y. Li, S. Zeng, G. Wang, P. Zhang, P. Hui, and D. Jin, "Modeling Spatio-Temporal App Usage for a Large User Population," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 1, Mar. 2019.

[24] C. M. R. Institute, "C-RAN: The Road Towards Green RAN," CMRI White Paper, 2011.

[25] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5g networks: New paradigms, scenarios, and challenges," *IEEE Communications Magazine*, vol. 55, no. 4, pp. 54–61, 2017.

[26] Next Generation Mobile Networks (NGMN) Alliance, "Description of network slicing concept," NGMN White Paper, Feb. 2015.

[27] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, May 2017.

[28] N. Nikaein et al, "OpenAirInterface: A Flexible Platform for 5G Research," in *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, Chicago, IL, Oct 2014, p. 33–38.

[29] D. Bega, A. Banchs, M. Gramaglia, X. Costa-Pérez, and P. Rost, "CARES: Computation-Aware Scheduling in Virtualized Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 7993–8006, 2018.

[30] I. T. U. R. S. (ITU-R), "Minimum requirements related to technical performance for imt-2020 radio interface(s)," Technical Report, 2017.

[31] M. Odini, "OpenSource MANO," IEEE Softwarization: A Collection of Short Technical Articles, Jul. 2016. [Online]. Available: https://sdn.ieee.org/newsletter/july-2016/opensource-mano

[32] J. G. Herrera and J. F. Botero, "Resource Allocation in NFV: A Comprehensive Survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, Sep. 2016.

[33] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-QoS-Aware Fair Scheduling for LTE," in *Proceedings of the IEEE 73rd Vehicular Technology Conference (IEEE VTC 2011 Spring)*, Budapest, Hungary, May 2011.

[34] 3rd Generation Partnership Project (3GPP), "Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)," 3GPP Technical Report (TR) 45.820, Aug. 2015.

[35] X. Li, D. Li, J. Wan, A. V. Vasilakos, C. Lai, and S. Wang, "A review of industrial wireless networks in the context of Industry 4.0," *Wireless Networks*, vol. 23, no. 1, pp. 23–41, Jan. 2017.

[36] Google, "Google Project Fi," 2018. [Online]. Available: https://fi.google.com/about/

[37] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166–174, Oct. 2017.

[38] A. Ksentini and N. Nikaein, "Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[39] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN Slicing for a Flexible and Cost-Effective Multi-Service Mobile Network Architecture," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2017)*, Snowbird, UT, Oct. 2017.

[40] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[41] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*, Atlanta, GA, May 2017.

[42] S. K. Sharma, T. E. Bogale, L. B. Le, S. Chatzinotas, X. Wang, and B. Ottersten, "Dynamic Spectrum Sharing in 5G Wireless Networks With Full-Duplex Technology: Recent Advances and Research Challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 674–707, Feb. 2018.

[43] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 1, pp. 124–161, 2016.

[44] C. Williamson, E. Halepovic, Hongxia Sun, and Yujing Wu, "Characterization of CDMA2000 cellular data network traffic," in *The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05)l*, Nov 2005, pp. Z000–719.

[45] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *2011 Proceedings IEEE INFOCOM*, April 2011, pp. 882–890.

[46] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, April 2017.

[47] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (ACM IMC 2009)*, Chicago, IL, 2009, p. 267–279.

[48] M. Shafiq, L. Ji, A. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM 2012)*, Orlando, FL, Mar. 2012, pp. 1341–1349.

[49] M. Shafiq, L. Ji, A. Liu, and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," in *Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '11. San Jose, CA: Association for Computing Machinery, Jun. 2011, p. 305–316.

[50] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling Users in a 3G Network Using Hourglass Co-Clustering," in *Proceedings of the 16th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2010)*. Chicago, IL: Association for Computing Machinery, Sep. 2010, p. 341–352.

[51] H. Li, X. Lu, X. Liu, T. Xie, K. Bian, F. X. Lin, Q. Mei, and F. Feng, "Characterizing Smartphone Usage Patterns from Millions of Android Users," in *Proceedings of the 2015 ACM Internet Measurement Conference (ACM IMC 2015)*, Tokyo, Japan, 2015, p. 459–472.

[52] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting Whatsapp through Large-Scale Measurements in Mobile Networks," in *Proceedings of the 2014 ACM Conference on SIGCOMM (ACM SIGCOMM 2014)*, Chicago, IL, 2014, p. 133–134.

[53] Q. Deng, Z. Li, Q. Wu, C. Xu, and G. Xie, "An empirical study of the WeChat mobile instant messaging service," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, May 2017, pp. 390–395.

[54] J. Erman, A. Gerber, K. K. Ramadrishnan, S. Sen, and O. Spatscheck, "Over the Top Video: The Gorilla in Cellular Networks," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (ACM IMC 2011)*, Berlin, Germany, 2011, p. 127–136.

[55] Z. Li, X. Wang, N. Huang, M. A. Kaafar, Z. Li, J. Zhou, G. Xie, and P. Steenkiste, "An Empirical Analysis of a Large-Scale Mobile Cloud Storage Service," in *Proceedings of the 2016 Internet Measurement Conference (IMC 2016)*. Santa Monica, CA: Association for Computing Machinery, 2016, p. 287–301.

[56] European Parliament and European Union Council, "Reglamento General de Protección de Datos (RGPD)," Boletín Oficial del Estado, May 2016.

[57] A. Swanson, "Tenant networks vs. provider networks in the private cloud context," OpenStack Foundation, 2016. [Online]. Available: https://superuser.openstack.org/articles/tenant-networks-vs-provider-networks-in-the-private-cloud-context/

[58] Next Generation Mobile Networks (NGMN) Alliance, "5G Security Recommendations Package #2: Network Slicing (Version 1.0)," NGMN White Paper, Apr. 2016.

[59] 3rd Generation Partnership Project (3GPP), "Study on the security aspects of the next generation system," 3GPP Technical Report (TR) 33.899, Aug. 2017.

[60] 3rd Generation Partnership Project (3GPP), "Security architecture and procedures for 5G System," 3GPP Technical Report (TR) 33.501, Jun. 2018.

[61] 5th Generation Public Private Partnership (5G-PPP), "5G PPP Phase 1 Security Landscape," 5G-PPP Security Working Group White Paper, Jun. 2017.

[62] E. Dotaro, "5G Network Slicing and Security," IEEE Software Defined Networks, Jan. 2018. [Online]. Available: https://sdn.ieee.org/newsletter/january-2018/5g-network-slicing-and-security

[63] R. Singh, M. Fiore, M. Marina, A. Tarable, and A. Nordio, "Urban Vibes and Rural Charms: Analysis of Geographic Diversity in Mobile Service Usage at National Scale," in *The World Wide Web Conference (ACM WWW 2019)*, San Francisco, CA, 2019, p. 1724–1734.

[64] J. Q. Stewart, "A Measure of the Influence of a Population at a Distance," *Sociometry*, vol. 5, no. 1, pp. 63–71, Feb. 1942.

[65] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021," White Paper, 2017.

[66] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," *ACM SIGMOD 2016*, vol. 45, pp. 69–76, Jun. 2016.

[67] M. C. G.W. Milligan, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[68] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*. Portland, Oregon: AAAI Press, 1996.

[69] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, vol. 8, pp. 14–38, Oct. 1991.

[70] C. Stolojescu, I. Railean, S. Moga, and A. Isar, "Comparison of wavelet families with application to WiMAX traffic forecasting," in *12th International Conference on Optimization of Electrical and Electronic Equipment (IEEE OPTIM 2010)*, 06 2010, pp. 932–937.

[71] B. Cazelles *et al.*, "Wavelet analysis of ecological time series," *Oecologia*, vol. 156, no. 2, pp. 287–304, 2008. [Online]. Available: http://www.jstor.org/stable/40213251

[72] A. Roesch and H. Schmidbauer, "WaveletComp: Computational Wavelet Analysis. R package version 1.1," The Comprehensive R Archive Network (CRAN), Mar. 2018. [Online]. Available: https://CRAN.R-project.org/package=WaveletComp

[73] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 84–91, May 2016.

[74] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis, "Network Store: Exploring Slicing in Future 5G Networks," in *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture (ACM MobiArch 2015)*, Paris, France, Sep. 2015.

[75] I. F. Akyildiz, P. Wang, and S. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," *Computer Networks*, vol. 85, pp. 1–18, Jul. 2015.

[76] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks," in *Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies (ACM CoNEXT 2016)*, Irvine, CA, Dec. 2016.

[77] M. R. Sama, X. An, Q. Wei, and S. Beker, "Reshaping the mobile core network via function decomposition and network slicing for the 5G Era," in *Proceedings of the 2016 IEEE Wireless Communications and Networking Conference (IEEE WCNC 2016)*, Doha, Qatar, Apr. 2016.

[78] A. Mayoral, R. Vilalta, R. Casellas, R. Martinez, and R. Munoz, "Multi-tenant 5G Network Slicing Architecture with Dynamic Deployment of Virtualized Tenant Management and Orchestration (MANO) Instances," in *Proceedings of the 42nd European Conference and Exhibition on Optical Communication (ECOC 2016)*, Dusseldorf, Germany, Sep. 2016.

[79] 3rd Generation Partnership Project (3GPP), "NR and NG-RAN Overall Description, Stage-2 (Release 15)," 3GPP Technical Specification (TS) 38.300, Jan. 2018.

[80] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Perez, "Network slicing games: Enabling customization in multi-tenant networks," in *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*, Atlanta, GA, May 2017.

[81] Y. L. Lee, J. Loo, T. C. Chuah, and L. C. Wang, "Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[82] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez, "Optimising 5G infrastructure markets: The business of network slicing," in *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*, Atlanta, GA, May 2017.

[83] A. Okic *et al.*, "Analyzing Different Mobile Applications in Time and Space: a City-Wide Scenario," in *2019 IEEE Wireless Communications and Networking Conference (IEEE WCNC)*, Apr. 2019.

[84] J.X. Salvat *et al.*, "Overbooking Network Slices through Yield-driven End-to-End Orchestration," in *ACM 14th International Conference on emerging Networking EXperiments and Technologies (ACM CoNEXT 2018)*, Dec. 2018.

[85] T. L. Nguyen and A. Lebre, "Virtual Machine Boot Time Model," in *Proceedings of the 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP 2017)*, St. Petersburg, Russia, Mar. 2017.

[86] 5th Generation Public Private Partnership (5G-PPP), "View on 5G Architecture (version 2.0)," 5G-PPP Architecture Working Group White Paper, Dec. 2017.

[87] P. Sanders and C. Schulz, "Think Locally, Act Globally: Highly Balanced Graph Partitioning," in *Proceedings of the International Symposium Experimental Algorithms (SEA 2013)*, Rome, Italy, Jun. 2013.

[88] M. d. Berg, O. Cheong, M. v. Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Applications*, 3rd ed. Santa Clara, CA: Springer-Verlag TELOS, 2008.

[89] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "CloudIQ: a framework for processing base stations in a data center," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2012)*, Istanbul, Turkey, Aug. 2012.

[90] F. Z. Yousaf and T. Taleb, "Fine-grained resource-aware virtual network function management for 5G carrier cloud," *IEEE Network*, vol. 30, no. 2, pp. 110–115, Mar. 2016.

[91] 3rd Generation Partnership Project (3GPP), "Telecommunication management; Study on management and orchestration of network slicing for next generation network (Release 15)," 3GPP Technical Report (TR) 28.801, Jan. 2018.

[92] D. Telekom, "Deutsche telekom, niantic and mobiledgex partnership." [Online]. Available: http://bit.ly/dt_niantic

[93] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communications Surveys & Tutorials*, Mar. 2018.

[94] European Telecommunications Standards Institute (ETSI), "Improved operator experience through Experiential Networked Intelligence (ENI) Introduction - Benefits - Enablers - Challenges - Call for Action," ETSI White Paper No. 22, Oct. 2017.

[95] D. Bega *et al.*, "DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning," in *IEEE International Conference on Computer Communications (IEEE INFOCOM 2019)*, Paris, France, Apr. 2019.