Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

PhD Thesis

On the Effects of Resource Sharing on Mobile Network Deployment Decisions

Author:
Paolo Di Francesco

Supervisor:
Prof. Luiz A. DaSilva

9th May 2016

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed:

Paolo Di Francesco 9th May 2016

Summary

The explosion of data demand, market saturation, increasing competition and consequent rapid price drops in mobile communications contributed to lower profit margins for mobile network operators, forcing them to seek ways to decrease network costs. In this context, network resource sharing emerged as one of the most appealing strategies to substantially and sustainably decrease network costs. Different forms of network sharing have been proposed and standardized in the course of the years, from the simple sharing of antenna towers and masts to the sharing of all mobile network operators' assets. However, the extent to which these options can be exploited in practice is case specific and it ultimately depends on a number of technical, financial, marketing, and regulatory aspects. This thesis analyzes how network sharing impacts different aspects of mobile cellular networks, with a specific focus on network planning decisions.

By applying spatial analysis tools on large-scale mobile operator demand data, we first present evidence that operators' demand exhibits sufficiently low correlation in time and space, supporting the claim that network sharing can effectively improve the network efficiency of mobile operators. Then, by leveraging these data, we thoroughly investigate the *infrastructure sharing* effects on planning decisions, looking at two problems, specifically, *network consolidation* and *network evolution.* An important contribution of this study is that, for the first time, competition regulation constraints are modelled next to the traditional technical constraints, adding an extra dimension to the network planning problem. Indeed, our findings reveal that even when regulators enforce tight competition constraints infrastructure sharing still provides benefits to mobile operators.

The advancements in virtualization techniques applied to mobile networks are contributing to the emergence of new business models for the mobile communication market. As the relationship between over-the-top (OTT) service providers and mobile network operators (MNOs) is entwined more than ever, we investigate the factors that can re-shape this relationship. Through a quantitative analysis based on extensive simulations and regression analysis, we identify the factors, both technical and non-technical, that influence network planning decisions. Our findings highlight that infrastructure costs structure and the characteristics of the demand, especially its spatial distribu-

tion, are important; but again the role of regulators, in this case in the form of new bands made available for broadband mobile communications can play a fundamental role in defining Service Level Agreements between OTTs and MNOs.

Next generation mobile standards (e.g., LTE-Advanced) will be designed to support the aggregation of spectrum bands that are both contiguous and not-contiguous. The non-contiguous spectrum aggregation in particular seems designed to stimulate spectrum sharing strategies between mobile operators that have exclusive access rights in different frequency bands. We investigate the dynamics involved in inter-operator spectrum aggregation. We first obtain the conditions under which two operators should cooperate, and allow each other to access portion of their spectral resources. Then, we devise a Bayesian game model. In this framework each independent operator can decide whether or not accessing the spectral resources of another operator and allowing other operators to access its own resources can improve its performance, without full information about the other operator's demand structure.. We design a distributed algorithm to approach a neighborhood of the Bayesian Nash Equilibrium.

In this thesis we investigate the benefits and limitations of several aspects of resource sharing in mobile networks. We explore in particular the impact of the infrastructure sharing paradigm on MNOs' network planning strategies and the emergence of new business models where OTT services providers have a major role on network deployment decisions.

Acknowledgements

First and foremost I would like to thank my supervisor Prof. Luiz DaSilva, who gave me the chance of working with him. He supported and motivated me throughout these years with great patience and enthusiasm. He has always been an example to follow and I could not have been luckier to have him as my mentor.

I would like to thank Prof. Linda Doyle for creating the fantastic work environment that is CTVR. Her enthusiasm and hard work have always been contagious contributing to the growth of the people surrounding her. As CTVR evolves into CONNECT, I am sure she will be even more the engine of this fantastic group. I would also like to thank Mr. Robert Mourik and Mr. Eddie Gleeson from Telefónica for the inspiring conversations we had and for the data they kindly supplied to me on which this thesis is based on.

My deepest thank also goes to all the people have met and have been close to me during this long journey as only a PhD can be. To all the senior researchers, Seamas, Yong, Nick, Nicola, Hamed, Johann, Tim, Irene, Paul and many others, not only for being *the shoulders of the giants* over which we stand, but also for contributing to the spirit of CTVR. A special mention goes to Francesco who worked very closely with me and addressed all the possible questions I could even come up with (even the silly ones). He gave an incredible boost to my self-confidence.

To Danny for being such a genuine person, the time spent with him has always been *energetic*. To Francisco for trying always to challenge me in any possible way. To Carlo and Lele for reminding me everyday the pleasure of being Italian. To Jacek for the really good time we spent together as visiting students at Virginia Tech and Washington DC, and for providing constructive comments to my work any time I needed. And to all the guys in the PhD lab (some of which moved to the "PostDoc room"), Arman, Justin, Pedro, Jonathan, Elma, Eamonn, Ioannis, Diarmuid, Jasmina, and Uche with whom I shared so many great moments. Without them my time in Ireland would have been much harder. To the guys of the WTS team (special thanks to Andrei) and in general to anyone who accepted to play football with me at least once. I rarely had so much fun in my life.

My never-ending gratitude goes to my family, and in particular to my parents Angelo and Tina

for being always supportive and raising me in the way I am. They thought me many things, but most importantly, that the hard work at the end of the day always pays off. I hope I can be at least half the person they are.

Finally, and most importantly, my most sincere thanks go to Carolina. She helped me to stay focused from the beginning to the end of this journey with her advices and constant encouragements. She has simply brought me where I am today. Without her, none of this work would have been possible. I dedicate this thesis to her.

Contents

List of Figures

List of Tables

1 Introduction

NOWADAYS the ability to share resource is a vital component to the survival of mobile network operators. In fact almost on a daily basis we witness new sharing agreements throughout the world. Since sharing agreements are generally established on a long-term basis (generally years), it becomes clear that sharing plays an important role on the planning and designing of present and future mobile networks. Planning and designing a mobile wireless system to offer coverage and access services to customers is a complex task per se. Finding the appropriate mix of deployment strategy and technologies requires operators to consider not only technological aspects, but also regulatory and economical ones, as well as their interaction. Considering the ability to share on top of these aspects adds another dimension to the network planning problem.

In this thesis we investigate the benefits and limitations of resource sharing in mobile networks and we analyze the impact of the infrastructure sharing paradigm on how mobile network operators and service providers plan their present and future networks. We show that tight competition regulation somewhat limits, but it does not completely jeopardize, the gains introduced by sharing, while having the secondary effect of encouraging the wider deployment of next-generation technologies. We also assess which aspects, among technological and non-technological ones, contribute the most in shaping service-driven network planning decisions considering the entwined relationships among mobile network operators, service providers, and regulators.

1.1 Overview and Motivation

The past two decades have witnessed a world-wide explosion of mobile wireless connectivity. In many countries, the mobile market penetration has gone well above 100%, with the average for the EU countries at roughly 124%. In many ways this success was fostered by the conventional mobile market model, which was based on exclusive ownership of both the network infrastructure and frequency spectrum license by the mobile network operators (MNOs). The exclusive ownership model dates back to the 1990s with the abolition of monopolies, when the telecommunications market experienced a wave of important changes. The entry of new players stimulated competition on pricing, product differentiation, and technologies advancement amongst mobile network operators to attract and keep customers. However, because of the high cost barriers to entry in the mobile market, the market configuration evolved from a monopoly to a oligopoly.

The 2G era was driven by voice services. At that time, MNOs experienced a tight relation between traffic served and revenues, since subscribers paid for the usage of the resources in terms of voice minutes. Despite competition increase, MNOs continued to manage their networks as a monopolistic operator, i.e. they took charge of *all the aspects* of the value chain model using a vertically integrated approach.[1] This model required the MNOs to have strict control over the deployment and maintenance of the physical infrastructures, services and content delivery, marketing, and billing [1, 2] (see Fig. 1.1 *single ownership*).

The existing relationship between traffic and revenue started to change with the introduction of 3G technologies on top of the 2G ones in the early 2000s. Market saturation, increasing competition and consequent rapid price drops contributed to lower MNOs' profit margins, leaving no choice to the MNOs but to seek new market models. Consequently, MNOs were forced to consider new ways to improve the cost efficiency of their networks by, for instance, reconsidering the provision of coverage and capacity everywhere all by themselves. The introduction of 4G technologies in the early 2010s and the simultaneous success of internet based services such as Netflix, Spotify and Google made these changes even more pressing. Furthermore, with the proliferation of mobile devices such as smartphones, tablets, and other data-hungry devices, MNOs today face a challenge. The growing demand for bandwidth and capacity has prompted costly infrastructure enhancements; subscribers have grown accustomed to services like mobile streaming and mobile video uploading and they are unwilling to tolerate a fee increase. While network disruptions, as it was the case for AT&T in 2010 caused by the first iPhones [3], are unlikely to happen again, the need for increasing data-rate endangers the profitability of running a cellular network [4].

Network sharing emerged as one of the most appealing mechanisms to substantially and sustainably decrease network costs. Different forms of network sharing have been proposed and standardized [5] in the course of the years, from the simple sharing of antenna towers and masts to the sharing of all MNOs assets. Depending on the level of sharing, it is estimated that operators can reduce their costs by 20-50% [6]. However, the extent to which these options can be exploited in practice is of course case specific and it will ultimately depend on a number of technical, financial,

[1] The purpose of creating a value chain is to understand where the revenues are generated and the costs are incurred.

marketing, and regulatory aspects.

Resource sharing can be classified into two basic categories: (i) passive sharing, and (ii) active sharing [7]. As the name itself suggests, passive sharing refers to the sharing of passive infrastructure, such as sites and building premises. Passive sharing is a moderate form of network sharing where multiple networks share physical space. Active sharing instead is a more complex process where MNOs share elements of the radio access network (RAN), such as antennas, radio nodes, node controllers, backhaul and backbone transmission, as well as elements of the core network like switches (see Fig. 1.1 *passive/active sharing*). In addition, active sharing includes roaming, which allows an operator to use the network of another operator in places where it has no coverage or infrastructure of its own. The roaming-based sharing paradigm in mobile wireless networks has opened up new business opportunities for MNOs. One example are mobile virtual network operators (MVNOs) (see Fig. 1.1). MVNOs are wireless communication service providers that own neither the infrastructure (i.e. the radio access network) nor the spectrum. Usually, MVNOs enter into service level agreements with one MNO, owner of infrastructure and spectrum license, to obtain access to network services at wholesale prices and then reset their own retail prices independently [8, 9].

In the past decade, internet-based service providers have become another important stakeholder in the mobile network market. The internet based service providers that generate large amount of traffic (e.g. Netflix is responsible for one-third of the peak-time traffic in the United States) have enjoyed, to some extent, a *free ride* on any technological advancements since they do not contribute in any way to the deployment of the infrastructure to obtain connectivity. The deployment of the required infrastructure affects both cable network providers and MNOs. One solution proposed by some telecommunication providers is to supply *premium access* for such service providers in exchange for additional fees. This solution however implies a revision of the concept of *network neutrality* and it has led to a fierce discussion amongst authorities, telecommunication firms, and the public.

We have recently proposed a new vision of future networks, called *Network without Borders* (NwoB), based on a market-place of virtual network operators. It provides connectivity to the end user from a pool of shared resources (e.g., base stations, spectrum, backhaul, cloud resources etc.) [8, 10]. In this new extreme form of sharing, the resource can be supplied by the traditional mobile operators or it can be pooled by individuals that can provide access network infrastructure (see Fig. 1.1). This model can be enabled thanks to the introduction and advancements of virtualization techniques in wireless systems [11]. It extends the sharing economy deeper into the mobile network through extensively embracing the concept of sharing of all the resources.

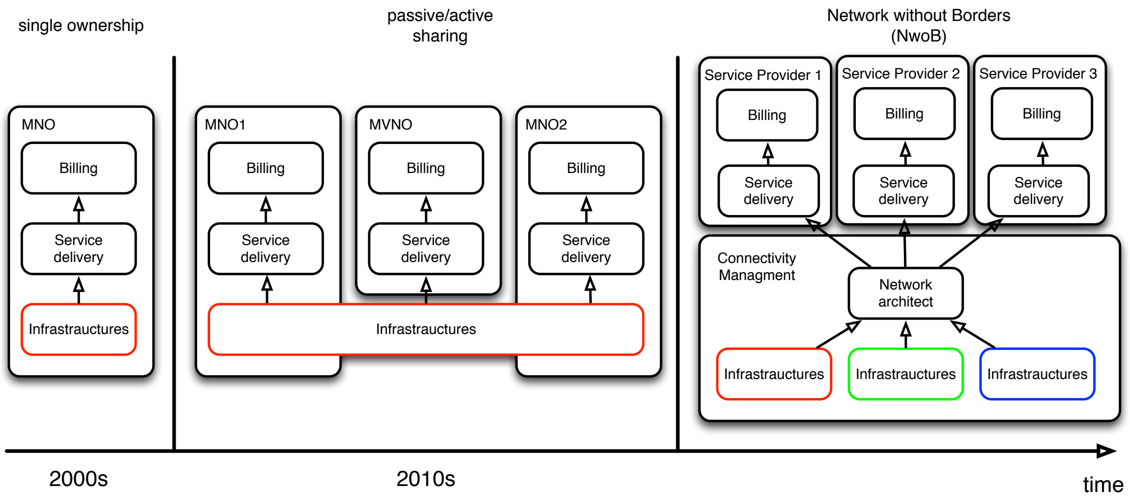


Figure 1.1: The evolution of mobile networks.

Providing cost effective and affordable mobile wireless services everywhere is one of the key requirements for future wireless systems success. However, achieving high data rates while keeping deployment costs low presents a number of technical challenges that have motivated and driven for many years research in the area of wireless systems and networking. The analysis in the context of mobile cellular communications includes several aspects: a) technical and cost performance, b) profitability and market position, c) conditions set by legacy systems, d) regulation and existing business models.

Both resource sharing and planning of mobile networks have received a great deal of attention in the research community. While efficient mobile network planning has been a widely studied research topic for decades, resource sharing is gaining its momentum in the past few years and it is still a hot topic. However, little attention has been given to the two problems combined. When considering the way resource sharing and planning interact and affect each other, we face a whole new multi-dimensional problem that has not been properly explored yet by the research community. In this thesis we aim to fill this existing gap.

1.2 Focus

This thesis examines planning decisions in mobile networks. The discussion revolves around resource sharing approaches and cooperative strategies among actors in the mobile telecommunication market. We discuss several ways through which the transition towards NwoB can begin. We consider to what extent, and in what context mobile operators can, will, and should cooperate with each other, focusing on network planning decisions. New networks can be deployed through joint ventures between operators, and existing ones can be managed jointly. In this regard, we will consider how such cooperation impacts the planning and upgrading of networks already deployed, given that coverage and capacity constraints exist alongside with competition regulation ones.

Moreover, resource sharing and virtualization techniques work as enablers for revolutionary business models. In this regard, the role of over-the-top (OTT) service providers in the market affects the operators' ability to generate profits. We will explore the economic incentives that enable OTT service providers and MNOs to cooperate towards the deployment of *service-driven* mobile networks.

The mobile market is rapidly changing and becoming more complex and sharing will play an important role in it. Planning resources efficiently will always be a central topic in mobile networks thus it is of paramount importance to understand the dynamics involved between resource sharing, new business models, and network planning.

1.3 Key Contributions

As a result of our work, this thesis makes the following contributions.

1.3.1 A New way to Use Real Data for Planning Purposes: a Correlation Study and a Modelling Framework

The first step of this thesis is to assess the correlation in space and time of the traffic demand between mobile network operators. For this study, we make use of two datasets in the form of call detail records (CDRs) supplied by two Irish nation-wide operators collected at their core networks. Based on our analysis of this raw data we deliver a quantitative study on the benefit of sharing, providing evidence that, when the spatio-temporal correlation of the demand between two operators is low enough, sharing should be encouraged. In this study at the beginning of Chapter 3, for the first time, we are able to compare the variation in the traffic demand of two mobile network operators covering the same territory.

The second step in this contribution is to propose a methodology to build a modelling framework suitable to study the planning of mobile networks at a large-scale by combining data available from different sources. Some of the data used are not, for proprietary reasons, publicly available, while others are. The proprietary data are the ones previously described. The non-proprietary data concern operators' infrastructure location deployment and demographic information, both of which are publicly accessible. In a nutshell, we provide a tool for researchers interested in studying network planning in real settings at a large-scale with a relatively low complexity. The second part of Chapter 3 details the relevant information needed to build our large-scale modelling framework for the Irish case but also how the methodology employed can be easily extended to other countries. The non-proprietary information about this framework is publicly available and can be freely used and/or modified to support further research in this field. Most of the results obtained in this thesis are based on this framework.

1.3.2 Infrastructure Sharing and its Consequences on Network Consolidation and Network Evolution Planning

The second step of this thesis is to quantify the benefits of the cooperation among traditional MNOs through infrastructure sharing and to assess to what extent it impacts the way MNOs can upgrade, maintain, and decommission part of their network. The discussion of infrastructure sharing in this case is limited to the Radio Access Technology (RAT).

Chapter 4 can then be divided into two parts. The first part of Chapter 4 explores how, starting from an already existing deployment, network consolidation through infrastructure sharing can provide significant cost savings in terms of the number of base stations needed to satisfy coverage and capacity constraints. We also propose a methodology to select base stations to be decommissioned. The second part of Chapter 4 looks at a longer time frame. It investigates how sharing agreements between two operators and competition regulations, as expressed by a *local* version of the Herfindahl-Hirschman Index (HHI), affect the operators' decisions concerning the upgrading, maintaining, and decommissioning parts of their network. In this chapter we propose a family of algorithms to be employed to schedule the changes to a network in a cost efficient way while satisfying the demand and complying with regulatory constraints.

1.3.3 Sensitivity Analysis on Service-Driven Network Planning

The third step of this thesis looks at the inter-relationships between decisions taken by mobile network operators and over-the-top service providers and the shape of *service-driven* networks. In particular, through a multi-dimensional sensitivity analysis, we seek to understand which ones among a wide range of parameters under scrutiny are the most significant ones on three aspects: the definition of the Service Level Agreements (SLAs) between MNO and OTTs, and the subsequent decisions on new infrastructure deployment taken by the MNO and the OTTs.

The first contribution of Chapter 5 is to identify the set of parameters of interest and their range of variation. Our model includes technical aspects such as network capacity, coverage evaluation and non-technical aspects, such as cost of deploying certain technologies. The second contribution is the outcome of the sensitivity analysis that captures and reveals which of those parameters will have the strongest impact on the emergence of service-driven networks.

1.3.4 Spectrum Aggregation

The last step of this thesis examines spectrum sharing from the perspective of spectrum aggregation. We extend the notion of Carrier Aggregation (CA) by exploring the possibility to aggregate non-contiguous portions of licensed spectrum that belong to different mobile operators.

The first contribution of Chapter 6 is to derive the conditions under which mobile operators allows other operators to access their spectral resources. Subsequently, we propose a Bayesian game-based framework. In this framework each independent operator can decide whether or not dynamically aggregate resources from other operators and allowing other operators to access its own resources can improve its performance without full information about the other operators. Then, we design a distributed algorithm to approach a neighborhood of the Bayesian Nash Equilibrium.

1.4 Publication

In the following, we detail the contribution in terms of dissemination of the work. The publications directly related to this research are: 1, 2, 3, 5, 6, 8, 10. The remaining papers were published in the course of my PhD studies as a result of collaborations in other projects.

Journals:

1. P. Di Francesco, J. Kibiłda, F. Malandrino, N. Kaminski, L. A. DaSilva, “Sensitivity Analysis on Service-Driven Network Planning”, currently under review.
2. P. Di Francesco, F. Malandrino, L. A. DaSilva “Assembling and Using a Cellular Dataset for Mobile Network Analysis and Planning”, currently under review.
3. P. Di Francesco, F. Malandrino, T. K. Forde, L. A. DaSilva “A Sharing- and Competition-Aware Framework for Cellular Network Evolution Planning”, *IEEE Transactions on Cognitive Communication and Networking*, September 2015.
4. P. Di Francesco, S. McGettrick, U. K. Anyanwu, J. C. O′Sullivan, A. B. MacKenzie, L. A. DaSilva “A Split MAC Approach for SDR Platforms” in *IEEE Transactions on Computers*, February 2014. [ToC2014]

Conferences:

5. J. Kibiłda, P. Di Francesco, F. Malandrino, L. A. DaSilva, “Infrastructure and Spectrum Sharing Trade-offs in Mobile Networks”, *IEEE Symposia on New Frontiers on Dynamic Spectrum Access Networks (DySPAN)*, October 2015.
6. J. Tallon, A. Pushmann, F. Paisana, J van de Belt, P. Di Francesco, N. Kaminski, H. Ahmadi “Coexistence Through Frequency Decimation and Markov Chains”, in *IEEE Symposia on New Frontiers on Dynamic Spectrum Access Networks (DySPAN)*, October 2015.
7. P. Di Francesco, F. Malandrino, L. A. DaSilva “Mobile Network Sharing Between Operators: A Demand Trace-Driven Study” in *ACM SIGCOMM Capacity Sharing Workshop (CSWS)*, August 2014. [csws2014]
8. L. A. DaSilva, J. Kibiłda, P. Di Francesco, T. Forde, L. Doyle “Customized Services over Virtual Wireless Networks: The Path towards Networks without Borders” in *Future Network and MobileSummit 2013*, July 2013. [fnms2013]
9. Y. Xiao, C. Yuen, P. Di Francesco, L. A. DaSilva “Dynamic Spectrum Scheduling for Carrier Aggregation: A Game Theoretic Approach”, in *IEEE International Conference on Communications (ICC)*, June 2013. [icc2013]
10. A. Pushmann, P. Di Francesco, M. A. Kalil, L. A. DaSilva, A. Mitschele-Thiel “Enhancing the Performance of Random Access MAC Protocols for low-cost SDRs” in 8th *ACM International Workshop on Wireless network testbeds, experimental evaluation and characterization (WiNTECH)*, September 2013. [wintech2013]
11. P. Di Francesco, S. McGettrick, U. K. Anyanwu, J. C. O′Sullivan, A. B. MacKenzie, L. A. DaSilva “A Split Architecture for Random Access MAC for SDR Platforms”, in 8th *Conference on Cognitive Radio Oriented Wireless Networks (CROWNCOM)*, July 2013.
12. J. C. O′Sullivan, P. Di Francesco, U. K. Anyanwu, L. A. DaSilva, A. B. MacKenzie “Multi-hop MAC Implementations for Affordable SDR Hardware”, in *IEEE Symposia on New Frontiers on Dynamic Spectrum Access Networks (DySPAN)*, May 2011. [dyspan2011]

2 Background and Related Works

This chapter reviews the literature directly related to the work conducted in this thesis. These works relate to three main concepts: resource sharing, network planning optimization, and the use of real data (i.e., deployment data, traffic demand, and demographic data) to analyze mobile networks. While extensive literature on resource sharing and network planning exists, real data to study cellular networks have been seldom used due to limited availability, especially concerning data traffic demand. To the best of our knowledge, we are the first to propose a study that considers resource sharing, network planning optimization, and real data in combination. In the following we present each topic separately, discussing studies that include in their analysis more than one of these aspects in combination.

2.1 Resource Sharing

Mobile network resource sharing can be classified as either *passive* or *active* [6, 7]. Passive sharing takes place between mobile operators that decide to share base station sites and their basic installations, such as mast, cooling equipment and power supply. Active sharing, on the other hand, involves some level of abstraction (virtualization) of the shared physical resources. These resources include physical infrastructure elements such as base stations, or time-frequency spectrum blocks. In the first instance, the resources shared are known as *infrastructure sharing*, while in the second they are tied to the notion of *spectrum sharing.* In Fig. 2.1 we illustrate the conceptual differences between *infrastructure sharing* and *spectrum sharing.*

2.1.1 Infrastructure sharing

Both industry and research communities have recognized the importance of network infrastructure sharing in the evolution of mobile networks. For example, 3GPP standardization efforts were initially defined in two documents, one describing the service aspects and requirements for network

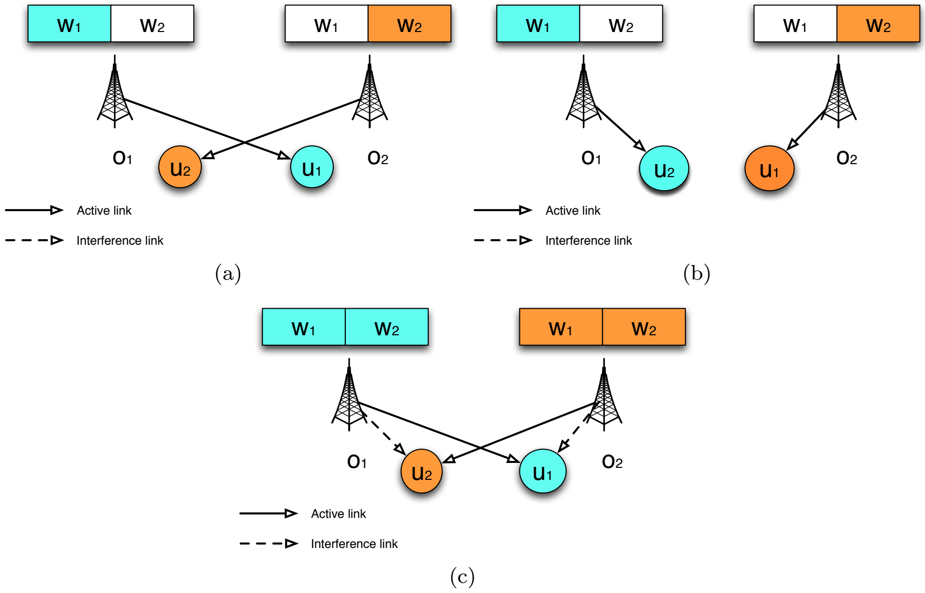


Figure 2.1: High level view of resource sharing. Two operators (1 and 2) with subscribers (u_1, u_2) (a) exclusively using their infrastructure (o_1, o_2) and spectrum (w_1, w_2), (b) exclusively using their spectrum (w_1, w_2), yet, using shared infrastructure (o_1, o_2) and (c) using shared spectrum (w_1, w_2), yet, not sharing their infrastructure (o_1, o_2).

sharing [12] and the other defining architectural changes and functions necessary to allow different core network operators to share a single radio access network [5] in order to meet the requirements set in [12]. In [5] two approaches are set to share the radio access network as depicted in Fig. 2.2: (i) the Multi-Operator Core Network (MOCN) and (ii) the Gateway Core Network (GWCN). In the former, only the radio access network part is shared. Maintaining a strict separation between the core network and the radio network has a number of benefits related to service differentiation and interworking with legacy networks. In the latter, instead, besides the radio access, some parts of the core network architecture are also shared, e.g., the Mobility Management Entity (MME), enabling additional cost savings compared to the MOCN.

As these 3GPP standardization efforts continued, more complex scenarios for cooperation have been added to complement already existing sharing capabilities as illustrated in [13]. Throughout Europe, MNOs have also increasingly employed RAN sharing, and in some cases this strategy has been brought to the extreme by completely merging two networks. This is the case for Newco, born from the fusion between Telia and Telenor in Denmark [14], and the Net4Mobility joint venture between Tele2 and Telenor in Sweden.

Existing works on infrastructure sharing problems follow two lines of research: (i) works focused on techno-economic modelling of network sharing and (ii) works focused on evaluating practical resource management techniques to evaluate the effectiveness of network sharing.

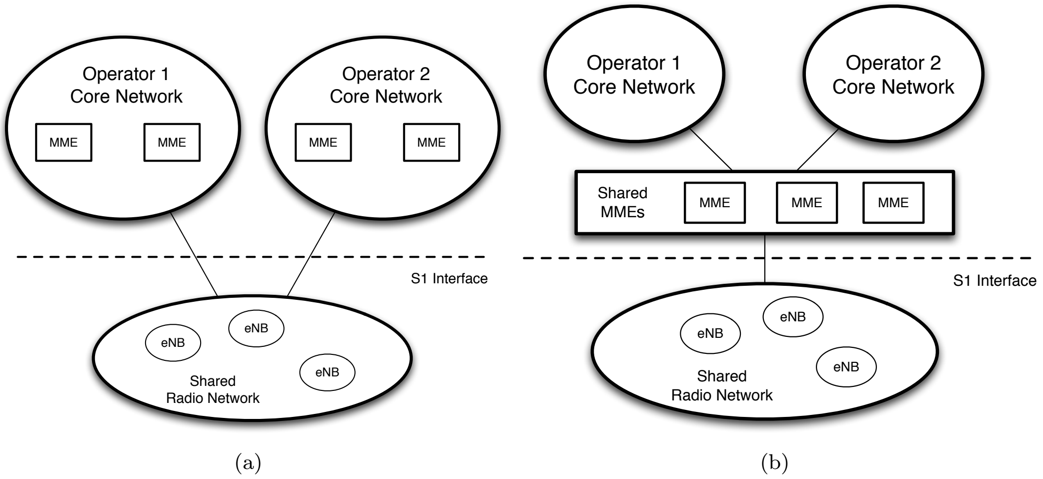


Figure 2.2: (a) MOCN architecture and (b) GWCN architecture.

The first category includes studies, both qualitative and quantitative, of different sharing scenarios, mostly modelling capital expenditures (CAPEX) and operational expenditures (OPEX). For example Markendahl in [15] analysed cooperative sharing arrangements highlighting the diversity of approaches currently being used in existing networks, their successes and failures in different settings. A study of Pakistan's experience of network sharing, presented by Choudhary et al. in [16], indicated the varying economic gains made in a still-developing market by adopting different sharing strategies. Overall these studies pointed to a process of learning within the industry as to which sharing modes allow for both competition and cooperative sharing to succeed.

The second category instead focuses on the practical side of infrastructure sharing. For example, Panchal et al. in [17] performed a simulation study that demonstrated how infrastructure sharing is the most effective form of sharing in terms of user performance. However, the study relied on highly idealized topologies, i.e., a hexagonal grid model both with co-located and not co-located base stations, and a simple on/off traffic generator, an approach often taken in these types of studies. Hua et al. in [18] investigated the benefits of cooperation among cellular operators using stochastic geometry, deriving average data rate and throughput under different sharing strategies. They showed that between 30% to 120% gains in (average) data rates per user can be achieved. While in this thesis we are not interested in the "green networking" aspect of infrastructure sharing that essentially consists in switching off and on base stations following daily traffic fluctuation, as presented in [19, 20], there is some potential overlap with our network design optimization purpose. Indeed we recognize that the solutions to the problems of infrastructure sharing for network planning and infrastructure sharing for energy savings are orthogonal, altogether compatible and to some extent, complementary.

Regulation on Network Sharing

Activities involving sharing agreements need antitrust authorities' approval. Regulators have generally reacted positively towards infrastructure sharing since they acknowledge its potential positive impact on the efficiency of utilization of national spectral resource. However, they must evaluate these positive aspects against possible competition concerns arising from a decreasing *network competition*. The major challenge faced by regulators is to distinguish between cases where dominant operators exploit infrastructure sharing to harm competition from cases where non-dominant operators need to stipulate sharing agreements, whether passive or active, to sustain a healthy competitive market. These competitive issues are usually undertaken on the basis of local competition laws and typically assess whether (i) the network efficiency gains outweigh any competitive harm and (ii) whether the same level of efficiency can be achieved in a less harmful manner. Assessment of both kinds is rather complex since it depends on the time frame considered. In the short term, regulatory measures aiming to boost competition may harm competition in the longer run. For example, regulatory approval (or mandate) of shared access to infrastructure and facilities of a competitor tends to increase competition in the short term. However, it may reduce competition in the long run as it decreases the incentives for newer generation network roll-out hence decreasing the likelihood of two or more networks to compete in the long term [21]. In light of these issues, regulators must consider both retail and wholesale mobile markets. For example, where there is effective end-to-end competition in the retail market, it is usually not necessary to regulate the wholesale market [21, 22].[1] However, as suggested by Beckman and Smith in [23], with the appropriate regulatory framework, the benefits of network sharing far outweigh the potential market issues. For example, Hultell et al. in [24], in order to preserve competition and reduce exposure to such risks, proposed that radio resource management (RRM) functions could be handled by third-party providers.

A common regulatory tool to measure the level of competition and market concentration is the Herfindahl–Hirschman Index (HHI) [25, 26]. It is given by the sum of the squares of shares held by each operator in the market, and takes values between 0 (a multitude of operators with a zero-share) and 1 (a monopolist with a 100% share). The HHI can be used to assess concentration in different aspects of the market, such as, among the others, overall market share, concentration in ownership of spectrum and concentration in ownership of network infrastructure. In this thesis, we propose a *local* version of the HHI in our regulatory framework that will be presented in details in Chapter 4.

[1] See [21] for the actions taken by national regulators in several European countries.

2.1.2 Spectrum sharing

Arguably spectrum is the most fundamental resource in enabling the success of any wireless communication system. MNOs have relied on the exclusive license assignment scheme in order to obtain the required bandwidth to ensure high quality of service and reliability for their subscribers. The exclusive license assignment scheme is implemented through public auctions regulated internationally by the International Telecommunication Unit (ITU) and nationally by local national regulatory agencies, e.g., Federal Communications Commission (FCC) in the US, or Office of Communications (OfCom) in the UK. Over many years, exclusive license policy has dominated spectrum assignment, but in the last few years it has started to be challenged.

The research community has tried to break down the view of the spectrum as an exclusive commodity, generating a whole line of research commonly referred to as *dynamic spectrum access* (DSA) [27, 28]. Currently, various network sharing models are being evaluated. Both the European Commission in [29] and the US President's Council of Advisors on Science and Technology (PCAST) in [30] have acknowledged the need for more elastic use of spectrum, promoting collective use and shared use of spectrum to exploit underutilized bands. Consequently, we have seen a recent ruling by the FCC to open military frequencies in the 3550-3700 MHz (3.5 GHz) band to mobile broadband services. Some of these rules specify the protection of incumbent radar systems from interference and are managed using database-aided spectrum access as proposed for TV whitespaces [31]. However the questions of which technologies, e.g., LTE Assisted Access (LAA) and WiFi, will be allowed to exploit the new available unlicensed bands is far from settled.

Extensive research exists on spectrum sharing among operators. Jorswieck et al. in [32] evaluated gains in spectral efficiency for two spectrum sharing regimes: orthogonal spectrum sharing, where operators exclusively assign spectral resources coming from a shared pool, and non-orthogonal spectrum sharing, where the operators aggressively re-use spectrum by allowing more than a single operator to use a specific spectral resource in any given area at a particular time. This study [32] demonstrated that if frequency bands are allocated dynamically and exclusively to one operator (i.e. orthogonal spectrum sharing) the gains attainable can go from 50% to 100% in terms of achievable throughput. Moreover, if frequency bands are allocated simultaneously to two operators (i.e. non-orthogonal spectrum sharing), the gains can go even further.

The major challenge faced in spectrum sharing regimes is how to address the interference issue. Bennis et al. in [33] explored the spectrum sharing problem when operators coexist in the same frequency band. The outcome of the problem, when formulated as non-cooperative game, leads to a Nash Equilibrium (NE) that is non-efficient and non pareto-optimal. When re-formulated as a Stackelberg game (i.e., leader-follower approach), the solution led to a pareto-optimal outcome. Kang et al. in [34] provided a power control game with the objective of minimizing mutual interference. They showed that in 80% of cases cooperation performed better than when players acted selfishly. The authors noted that as the network size increases, the incentive to cooperate decreases drastically due to the increased interference; however, they pointed out that only marginal network separation is necessary to decouple the interference.

European funded projects (FP7) such as SAPHYRE [35], and SAMURAI [36] looked at spectrum utilization where cooperation, competition and aggregation are key characteristics. In particular SAMURAI investigated the ability to aggregate different component carriers as specified in LTE-Advanced, Release 10 [37], systems, although it only examined the intra-operator mechanisms. In this thesis, instead, we are interested in extending the notion of carrier aggregation by allowing dynamic inter-operator aggregation as a mean for spectrum sharing.

As spectrum usage becomes more fluid, the concept of *fungibility* of spectrum becomes an increasingly important issue [38]. Fungibility is a term, originating from economics, indicating whether two units of a certain good can be seen as interchangeable. Oil is fungible in that, at least in a first approximation, a barrel of oil has the same utility and the same value as any other barrel of oil. Books, on the other hand, are not fungible – different books, and even different editions of the same book, are not interchangeable. The increase in spectrum trading resulting from the ability and motivation to rapidly change operating frequencies makes fungibility an important consideration for future radio systems. Network operators now have the ability to divert traffic between a number of different bands and technologies, buying this capacity on demand; fungibility provides an important tool in assessing the relative usefulness of these options based on the goals of the operator. In pursuit of refinement of the fungibility concept, Weiss et al. in [38] proposed several methods for calculating a score to quantify the relative fungibility of frequency bands. The scores provided by this work can be broadly categorized into probabilistic scores and distance based scores. In a follow-up work [39], Gomez and Weiss refined the determination of fungibility scores. Specifically, fungibility scores work focused on the subjects of coverage and capacity rather than the various dimensions presented in the original work. The coverage aspect was similar to the spatial dimension presented in the prior work, with the difference that cell radius for a given transmitter power and receive power were compared. A standard link budget calculation was used to determine the maximum accepted path loss at each frequency which was then used to determine the cell radius for each frequency. The capacity coverage sub-score compared the Shannon capacity at a given distance from the transmitter for each band. This sub-score captured a concept similar to that of the temporal dimension of the prior work, with the additional benefit of allowing for the comparison of varying bandwidths. Together these refinements clarified the determination of a fungibility score for two potential bands and focus the resulting value on the topics of coverage and capacity which are often the most important aspects when considering options for frequencies in network planning.

2.1.3 Virtualization and NwoB

Resource sharing is a key building block for virtualizing future mobile networks. Authors in [11, 40] provided a comprehensive survey of the radio access network (RAN) sharing functionality currently standardized and discussed in 3GPP [5]. The emergence of virtualization techniques in the wireless world is quickly gaining attention due to the ability to abstract, slice, isolate and share physical wireless network infrastructure and physical radio resources, where, ideally, a slice is a complete wireless virtual network [11].

We introduced in [10] Network without Borders (NwoB), a new concept of wireless networks, characterized by an extreme sharing regime. Operators construct their networks in a service-oriented fashion, exchanging resources from a shared pool sourced by network infrastructure providers as well as crowd-sourced from individuals, through a virtual marketplace. As described by Doyle et al. in a follow-up work [8], this new vision also entails a business paradigm shift with mobile network operators (MNOs) and mobile virtual network operators (MVNOs) having their role completely re-defined. The MVNOs can be now seen as specialized service providers who can help MNOs to attract more subscribers, while MNOs can produce more revenue by leasing slices to MVNOs. For instance, an over-the-top (OTT) service provider requesting a slice from an MNO means that the OTT wants to control a complete virtual network, from core network (CN) to air interface. Depending on the service offered, OTTs might need to ensure a minimum quality of service (QoS) to deliver their contents (e.g., HD video streaming). In this new business model, in exchange for a fee, the MNO would supply as many customizable resources as necessary to satisfy the QoS required in the areas served by the OTT. To meet the expected requirements, OTTs might decide either to enter into a service level agreement (SLA) with MNOs or to deploy their own network infrastructure or a combination of both. For example, the recently unveiled Google's Project Fi [41] offers to its subscribers both Wi-Fi, as part of Google's effort to deploy its own infrastructure, and LTE connection, as part of Google's MVNO agreement with traditional MNOs (i.e., Sprint and T-Mobile in the US). If we look at the business motivations behind Project Fi, we may see a strategy to pressure the bigger carriers into acting in a way to promote cheaper data plans, seamless WiFi transitions, that in return would create more internet access benefiting the actual service-oriented nature of Google [42]. Other similar examples include the Facebook Zero initiative [43] and the Twitter deals [44]. In other words, we may be entering the age of *service-driven network expansion*, and, more forward-looking, *service-driven networks*.

2.2 Network Optimization

Cellular network planning is a research topic that has been studied since the roll-out of the first commercial network. As illustrated in [45, Chapter 14], network optimization methods that use accurate performance evaluation, usually through Monte-Carlo simulations, provide detailed statistics, but they also require large computational effort especially if many network parameters are involved in the optimization (e.g., site location, pilot power assignment, antenna tilt, etc.). Differently, simplifying the system such that the remaining abstract model is easy to handle, to understand, and to evaluate, provides a way to evaluate network optimization problems more efficiently. If this is the case, the network optimization problems can be formulated as *mixed integer programming* (MIP) models [46]. Despite the abstraction, large instances of this class of problems are still too complex to be solved to the optimal solution within reasonable time. Often *heuristics* methods are used, e.g., simulated annealing, Tabu Search or greedy algorithms. The research following this line is vast and includes various aspects. For example, in [47–53] the authors studied the optimization of base station location in an area of interest. Some of these works [47–50] dealt with 3G systems and they were based on meta-heuristics aiming at minimizing the number of base stations to be deployed. Other more recent works [51–53] have focused on the same objective using similar approaches but on LTE networks. The work in [52] in particular used a model based on stochastic geometry, and the coverage probability as the metric to optimize.

Sometimes, if the interference modelling is kept simple and only few parameters are optimized, e.g., base station locations, network planning problems can be solved using *integer programming* even though they still suffer of high complexity for large instances. This is for example the case for Boiardi et al. in [54], where the authors studied the cost and energy savings of cellular networks in a small setting. They relied on *coverage points*, placed on a regular grid, and *traffic points*, placed randomly and associated to a uniformly-distributed demand using hard capacity constraints. Another way to model the interference is by generating an *interference matrix* where each element represents the fraction of service loss due to the presence of the considered interferent [55]. In this way, by approximating the impact of the interference, the authors were able to linearize the capacity constraints while still considering the interference.

Rarely network optimization problems are considered in combination with resource sharing. To date, and to the best of our knowledge, only Kibiłda et al. [56] and Cano et al. in [57] investigated network optimization planning related problems with infrastructure sharing. Kibiłda et al. [56] investigated the coverage efficiency obtained by combining existing real cellular networks exploiting the coverage redundancy in real deployments. In [57] the authors formulated the problem of whether MNOs with consolidated networks should strategically deploy additional LTE small-cells on their own or deploy shared LTE small-cells, hence also sharing the costs of the deployment.

The goal of this thesis is to develop an optimization method for radio network planning focusing on the installation of base stations. Hence, our objective is to design a network that is capable of supporting and satisfying coverage and capacity constraints at the minimum cost considering the possibility to share infrastructure, as we present in Chapter 4. Moreover, we consider the impact of several parameters on the planning of *service-oriented* networks analyzing the extent to which over-the-top service providers decide to rely on shared infrastructure or to deploy their own infrastructure, as we present in Chapter 5.

2.2.1 Cost Modelling

An important aspect of network optimization is the modelling of the costs associated with running a mobile network. The cost structure of building and running a mobile network include the costs, among the others, of: base stations site acquisition/lease, radio equipment, cooling systems, backhaul, and so forth. Several cost models have been proposed in the literature such as the one in [58] where Johansson provided detailed estimated figures breaking down the CAPEX and OPEX for heterogeneous infrastructure deployment, obtained by published reports (see references therein) and author's assumptions. Figures can also be obtained from corporate reports; however the figures are quite varying, see, for example, [59] and [60]. As Markenhal [1] has pointed out, the solutions and deployment strategies depend on the cost modelling assumptions. In Chapter 5 we perform a sensitivity analysis on different cost structures for several technologies.

Mölleryd and Zander in [61] proposed a cost model for spectrum. In particular, by assuming the price paid by the operators during the auctions as the marginal value of the spectrum, they analyzed the *engineering value* of the spectrum, calculated on the basis of comparisons of different network deployment options using different amounts of spectrum. Essentially, the engineering value of the spectrum can be expressed by the cost savings in the infrastructure of an operator's network obtained by having access to additional spectrum [61]. Han et al. in [62] built on the considerations in [61] and proposed a formal definition of the engineering value of the spectrum as well as the economic value of the spectrum.

While these papers addressed some of the economic effects of sharing, none of these has dealt with regulatory concerns regarding the ensuing market concentration which occurs through network sharing in mature mobile markets.

2.3 Real Data

In this thesis, we rely on three types of data: deployment data, demographic data, and traffic data. In this section we present the works that have used similar data and how such data have been exploited in the context of mobile networks. We give more attention to the ones that focused on either network planning or infrastructure sharing. Our contribution in this regard is in depicting a methodology that uses real data and that is tailored to network planning and in particular to infrastructure sharing.

2.3.1 Demographic Data

Demographic data have been used mostly in conjunction with network deployment problems. For example, Michalopoulou et al. in [63] study the interaction between cellular deployment and population density using tools from spatial statistics and spatial point processes. They take Germany as a case study, and charactarize infrastructure deployment as a spatial point process, depending on local population density. Demographic data describing population densities have always been used by operators, especially during the roll-out phase of their networks. Indeed, prior works such as [64] show that demographic data can be used to roughly estimate the spatial traffic demand assuming that at higher population densities correspond to higher traffic demand that has to be served by the network operators.

2.3.2 Deployment Data

Large-scale wireless networks are typically evaluated on standardized, synthetic topologies [65]. These synthetic topologies often feature regular, lattice-like, single-operator deployments. Unfortunately they are inappropriate to study mobile networks and network sharing given that the distribution in space of the base stations is an important factor and it has a significant impact on several properties of the network itself such as coverage and capacity [66]. Stochastic geometry applied to cellular communications has tried to fill in this gap by using *random spatial models* for the geographic distribution of the base stations. Assumptions of uniform distribution of base stations in a geographic area of interest, however, may be unrealistic for areas that are large enough to have a very skewed distribution of the population. One way to overcome this limitation is to characterize the relationship obviously existing between deployment and population density [63].

Another option is for studies of large-scale wireless networks to rely on data for real topologies. In many cases these data are either freely available online ([67–70]) or they can be obtained directly from the operators. Follow-up works on stochastic geometry and real base station deployments, such as [71] by Kibiłda et al. [52] by Guo and Haenggi, revealed that the spatial structure of base station deployments can be fitted by using clustered point-processes with important implications for network design and theoretical evaluation of network performance. Very few attempts have been made to study resource sharing using real deployments ([18, 19, 56]). Moreover, in the existing studies, the scale of the analysis is often constrained to a very limited area, generally an urban area. The whole picture is somehow incomplete since mobile networks operators (MNOs) are forced to deploy their networks on a national scale (e.g. because of regulatory constraints) and, due to regulatory constraints, they cannot neglect coverage of rural areas, which usually are the ones with a lower Return of Investment (ROI). For this reason, our thesis aims to analyze nation-wide deployments where both urban and non-urban areas can be considered. As our study will show, rural areas are also where operators could obtain the greater cost savings by sharing their resources.

2.3.3 Traffic Demand Data

Traffic demand data are also a valuable source of information for network planning. Unfortunately it is difficult to find studies that use real operator-supplied data; moreover, all of them rely on dataset(s) from only one operator, which is insufficient to study network sharing. There are a number of studies in the literature that aimed at analysing traffic dynamics in cellular networks; they can be grouped into two categories: field measurements-based and large-scale dataset-based.

Field measurements-based

Field measurement studies have the advantage of capturing the actual channel occupancy [72–74]. These studies have been performed in normal typical weekdays as in [72, 74] or during events gathering large amounts of people as in [73] during match days of the World Cup 2006 in Germany. All these works offered interesting insights about the demand and its fluctuations as well as the behaviour of the network in cases of different loading conditions [73]. In fact, they all provided evidence that by exploiting the statistical multiplexed nature of the traffic demand, mobile dynamic spectrum access techniques can increase the spectral efficiency. However, their foremost limitation is the difficulty to provide concurrent measurements at more than a few locations.

Dataset-based

Occasionally, mobile operators disclose demand information to individual research groups in the form of datasets. These datasets come in the form of traces or, equivalently, call detail records (CDRs) of voice sessions and data sessions collected at the core network, more specifically, at the Serving GPRS support node (SGSN). They present structured information for each record, e.g., timestamp, duration (or amount of bytes downloaded and uploaded), base station where the call was started, base station where the call was ended, and in rare cases, TAC code and user identifier. In some other cases, flow-level information can be also available.

Researchers studied different aspects of mobile networks using demand data. We have identified the following: (i) characterization of the user behaviour and characterization of the application popularity, and (ii) characterization of the network behaviour. In studying the user behaviour, the main objective is to extract statistics to describe how subscribers use the network with obvious implications on billing and network planning. For example, it has been shown in [75] that the call arrival process can be well approximated with Poisson arrivals and exponential holding times. In [76], Keralapura et al. analyzed the browsing behavior of mobile users in an American 3G data network, by monitoring 24 hours of IP traffic. Some studies noted that different types of data demand tend to be spatially correlated [77, 78]. These particular studies were concerned with web services and smartphone apps only; however, similar conclusions appeared to apply to mobile demand more generally. Shafiq et al. in [79] operated on flow-level information sent to and from cellular devices. Their study reveals that various data applications are not equally popular across all cells, and that “the popularity of some applications is more skewed than others across cells”. Moreover, a few applications dominate others given their relative traffic volume, and applications can be grouped into *traffic profiles* that describe application usage distribution for any given cell. Paul et al. in [80] also looked at individual subscriber behavior and traffic patterns, studying a nation-wide 3G network at the base station level. Differently, Shafiq et al. in previous works investigated application popularity and clustering [81] and device utilization [82] in a cellular network, obtaining useful insights that can be leveraged to fine tune network parameter settings such as inactivity timers of radio resource control (RRC) and the QoS profile settings and the radio network controller (RNC) admission control procedure.

2.4 Summary

In this thesis we have received access to traffic demand datasets from two nationwide Irish mobile operators collected at their core networks. Thus, and for the first time, we are given the possibility to compare the actual traffic demand from two operators covering the same large territory. Having access to these data from more than one operator opens up great opportunities to study mobile networks in a completely new way by taking into account the actual correlations in space of the traffic demand. Our study is unique because it combines all the aforementioned aspects (real data analysis, spatial distribution of traffic, network sharing, and network planning), and it considers the impact of the limitations imposed by the regulators on the savings when two networks are managed in a shared fashion in a mature market. Moreover, we explore the incentives existing in *service-driven* planning and how the interaction between mobile operators and service providers is affected by technical and non-technical factors.

3 Research Design and Methodology

IN this chapter, we first present our quantitative study on traffic demand correlation between two operators, using real traffic and deployment data. After assessing evidences that resource sharing between mobile operators should be encouraged when the spatial and temporal correlation of the demand are low enough, we present a new methodology to build a modelling framework that exploits and combines: traffic demand data, deployment data, and demographic data. The general purpose of the resulting modelling framework is to study cellular planning; however, in our case, we extend its use to the planning of shared networks.

3.1 Mobile Network Operators' Traffic Demand Correlation

Resource sharing provides improved resource utilization efficiency by statistically multiplexing the resources. For this reason, mobile network operators are greatly interested in it. Sensible as it sounds, there are several issues that could undermine the practicality and effectiveness of network sharing. Some are related to commercial agreements or competition issues. Some others, instead, are technical: intuitively, network sharing makes sense if the networks being joined and their demand are *different* enough. Joining two networks with very similar deployments and very similar loads has no effect on their ability to accommodate the peak load, as illustrated in Fig. 3.1.

Load and deployment are the foremost aspects to account for in studying the potential effectiveness of network sharing. In this section, we leverage on the data from two real-world traces, provided by two Irish network operators. Such information allows us to assess the practicality and the potential performance benefits of sharing capacity through real data, without the need to rely on (potentially, oversimplified) models and (potentially, unrealistic) assumptions.

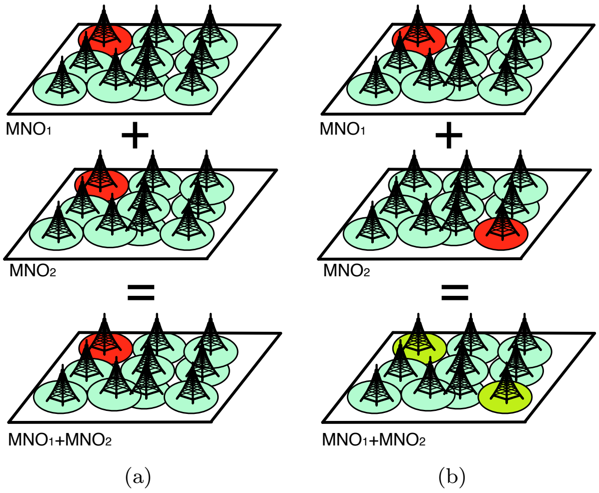


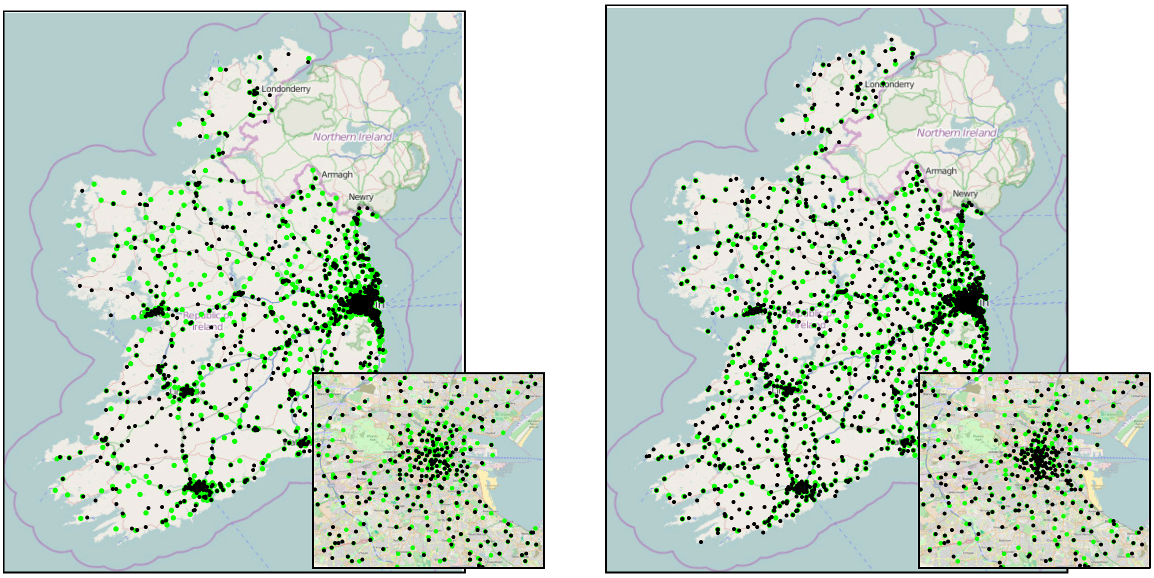
Figure 3.1: Network sharing: (a) combining two networks with very similar load patterns yields little or no benefit. Instead, (b) combining two networks with different load patterns results in a more evenly distributed load for both networks.

3.1.1 Datasets

Our datasets come from two Irish operators, Meteor and O2 (as of March 2015, Three). The datasets include a one week long call-detail record (CDR) information for both data and voice, concerning over 10,000 2G (i.e., GSM/GPRS) and 12,000 3G (i.e., W-CDMA/HSPA) transmitters distributed over the entire Republic of Ireland. For each transmitter, we know its position, azimuth and sectorization information, as well as its approximate coverage area in the form of power class. For each voice call and data session, we know the transmitter from which it is initiated and the transmitter at which it is terminated, the duration, and the amount of data transferred (e.g. in downlink and uplink). As was the case in [75], our datasets do not provide full information on the mobility of the users. Therefore, to circumvent this limitation, we assume that the entire call/data session took place entirely in the location of the initial transmission. Given the level of resolution we are interested in (e.g. hourly network usage) and the average short duration of calls and data sessions, this approximation does not affect our conclusions, as it has been shown in [75]. All the datasets are maintained in a MySQL server for processing purposes. Fig. 3.2 summarizes the nation-wide deployments for both GSM and 3G.

The locations of the transmitters grouped into base stations can also be found on the Irish regulator website [67]. Table 3.1 provides a summary of the information available online.[1]

[1] Last checked 1th June 2015.



(a)

(b)

Figure 3.2: (a) 3G and (b) GSM deployments. Dark points represent MNO$_1$ base stations; light green points MNO$_2$ base stations. The densely covered area in the East corresponds to Dublin (zoomed in the box).

Technology	MNO$_1$	MNO$_2$	Total
2G (GSM/GPRS)	1588	1166	2754
3G (W-CDMA/HSPA)	1209	1311	2520

Table 3.1: Number of base stations included in our datasets, for each operator and technology.

3.1.2 Global correlation

As discussed in Sec. 3.1 and summarized in Fig. 3.1, network sharing is ineffective when the networks being shared are too similar to each other. In this section, we analyse the correlation between the load of MNO$_1$ and MNO$_2$, in both space and time. A high degree of correlation would mean that the potential benefit of network sharing is limited; on the opposite, a lower degree of correlation would bode well for network sharing.

Time correlation

In our approach we represent the load of each sector (i.e., the area covered by each transmitter) of each operator through a time series. The time resolution is one hour. We consider as *load* the duration of voice calls and the amount of data exchanged in a data session. The traces do include the duration of data sessions, but such information is often unreliable, e.g., there are many hour-long sessions with no data exchanged.

The first aspect we study is the autocorrelation of the time series, shown in Fig. 3.3. The shape of all curves reflects well known daily patterns: there is high positive correlation at 24-hour intervals (and, to a decreasing degree, 48-, 72-, etc.), highly negative correlation at 12-hour intervals (and, decreasing in magnitude, at 36-, 60-, etc.). Similar effects were observed in [80]. Also notice how the two operators exhibit virtually the same behavior.

What is less expected and more interesting is the sharp difference between voice (Fig. 3.3(a)) and data (Fig. 3.3(b)), with the latter having a much lower correlation. Intuitively, data traffic tends to have a more irregular time evolution; this translates into a higher probability that different operators experience different load levels at a given time. This bodes well for the effectiveness of network sharing in current networks, where most of the load is due to data rather than voice, and even more so in future, with additional services such as gaming and tele-presence coming into play.

Fig. 3.3(b) indicates the difference between the busiest and median sectors: the correlation for the busiest sector is much higher. Intuitively, this suggests that the load of busy sectors follow very regular patterns, while less-used sectors have more changing loads. This may represent an issue, since busy sectors are exactly the ones that should benefit more from network sharing. We need a clearer view of how busy sectors are distributed in space, as described next.

Space correlation

Our purpose now is to understand how strong the space correlation of the demand is. In other words, if a sector is highly loaded, how likely is it that its neighboring sectors will also be highly loaded? Similarly to time correlation, space correlation is relevant to understand the effectiveness of network sharing: if busy (i.e., potentially overloaded) sectors come in large, compact clusters, then it is less likely that combining networks from different operators can do much about them.

Moran's index

Unlike for time correlation, there is no unique definition of space correlation. We employ Moran's index [83], also used in [80, 84] to study spatial aspects of network phenomena. In our context, we can define it as:

$$I_G = \frac{n}{S_0} \frac{\sum\limits_{i=1}^{n} \sum\limits_{j=1, j \neq i}^{n} w_{i,j}(x_i - \bar{X})(x_j - \bar{X})}{\sum\limits_{i=1}^{n} (x_i - \bar{X})^2},$$

where n is the number of sectors, x_i represents the load of sector i and $\bar{X}$ is the average load, and

$$S_0 = \sum_{i=i}^{n} \sum_{j=i,j\neq i}^{n} w_{i,j}.$$

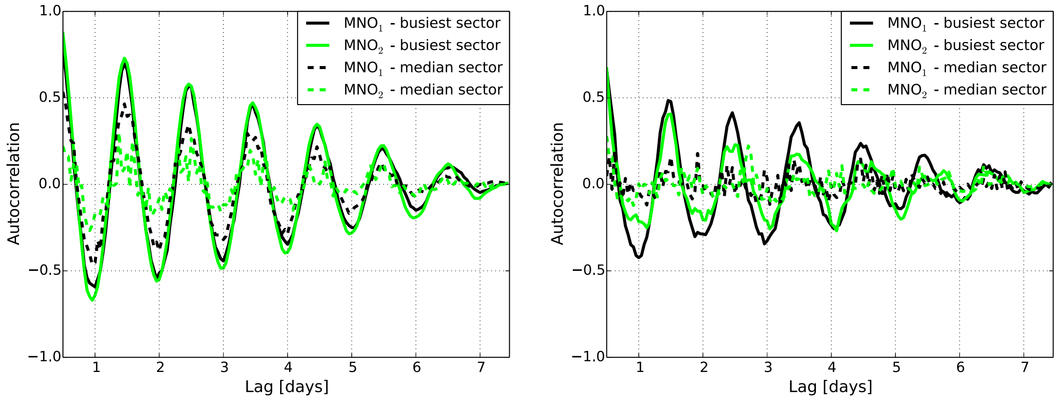
The weights $w_{i,j}$ represent in general the *distance weight* between two elements; usually, the Euclidean distance is used.

In network sharing scenarios, load can only be shared between *overlapping* sectors. Therefore, we adopt the following alternative definition of distance weight:

$$w_{i,j} = \frac{|A_i \cap A_j|}{|A_i \cup A_j|},$$

where A_i is the area covered by sector i. From our viewpoint, two sectors that do not overlap are infinitely distant from each other, as indeed there is nothing network sharing can do about their load.

The resulting correlation is plotted in Fig. 3.4. We can see that it is slightly higher during weekdays and during peak hours (around 8am and 6pm). However, the most important aspect to observe is that correlation levels are always very low.



(a) (b)

Figure 3.3: Autocorrelation for 3G (a) voice and (b) data.

Recall [83] that Moran's index is 0 for complete spatial randomness, 1 for perfect correlation, and −1 for perfect negative correlation. Our values seldom exceed 0.15, corresponding to positive but very weak correlation. This weak correlation in the spatial distribution of demand indicates that network sharing among operators is likely to be beneficial in handling instances of high demand.

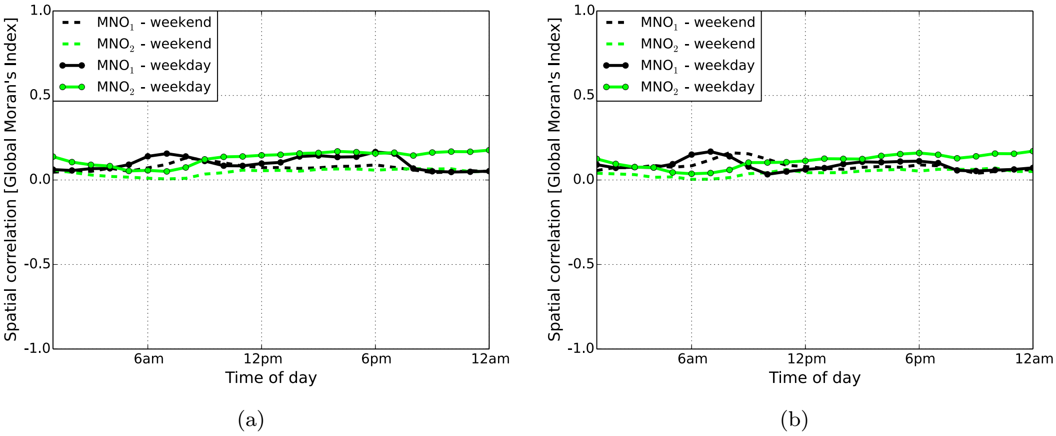


Figure 3.4: 3G data: space correlation (Moran's index) at different times of the day and for different spatial resolution: (a) Ireland and (b) Dublin.

3.1.3 The effectiveness of network sharing

So far, we have found several hints that network sharing is a promising way of tackling the load in cellular networks. Now, we want to go one step further, and assess *how much* we can actually gain from it. We start then by computing the local version of the Moran's index [85].

The index for sector i is defined as:

$$\frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{X}),$$

where

$$S_i^2 = \frac{\sum\limits_{j=1, j \neq i}^{n} (x_j - \bar{X})^2}{n-1} - \bar{X}^2.$$

Combining the index values for neighbouring sectors, we can divide them into four classes, namely:

HH high-load sectors surrounded by other high-load ones;

HL high-load sectors surrounded by low-load ones (*hot spots*);

LH low-load sectors surrounded by high-load ones (*cold spots*);

LL low-load sectors surrounded by other low-load ones.

Notice that the classification is made on a per-operator basis, i.e., we do not mingle together the traces of the two operators.

Indeed we are especially concerned with hot spots, i.e., sectors in class HL. These sectors could be linked to the so-called *flash crowds*, i.e., groups of people sharing the same location that suddenly become interested in downloading some data. Such events are often impossible to foresee, and represent a significant challenge for the operations of cellular networks [86].

Therefore, we look for HL sectors (hot spots), that overlap with sectors of the other operator that have low load, i.e., that are in LL or LH class – just like in the right-hand case described in Fig. 3.1. For these sectors, network sharing can better distribute the load, and thus improve the network performance.

Fig. 3.5 shows the number of hot spots when the MNO_1 and MNO_2 networks are operated separately or jointly, for different times of the day, during weekends and weekdays. The most important aspect to observe is the sharp decrease in the number of hot spots brought by network sharing. This holds for any time of the day, for both networks, for both weekdays and weekends: enabling network sharing consistently translates into fewer hot spots.

This is clearly very good news: as we mentioned, hot spots represent one of the most significant challenges that cellular networks have to face, and a technique as simple and cost-effective as network sharing proves very effective in curbing it.

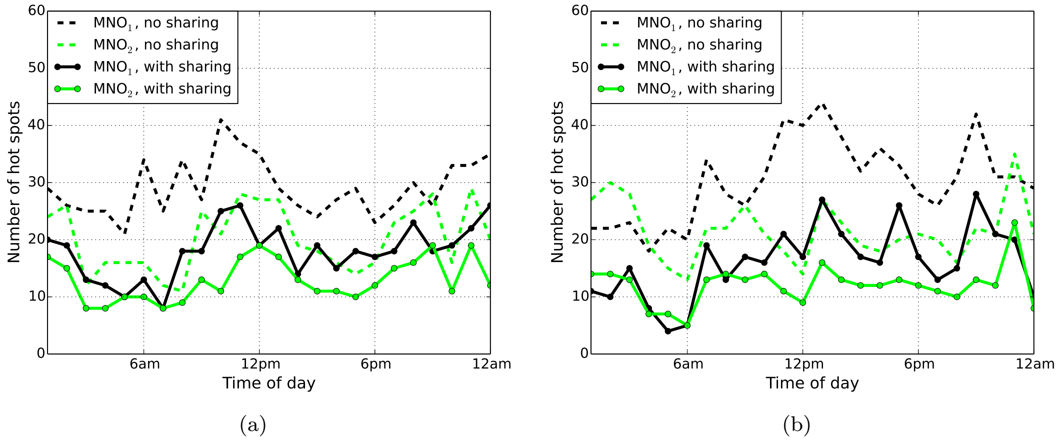
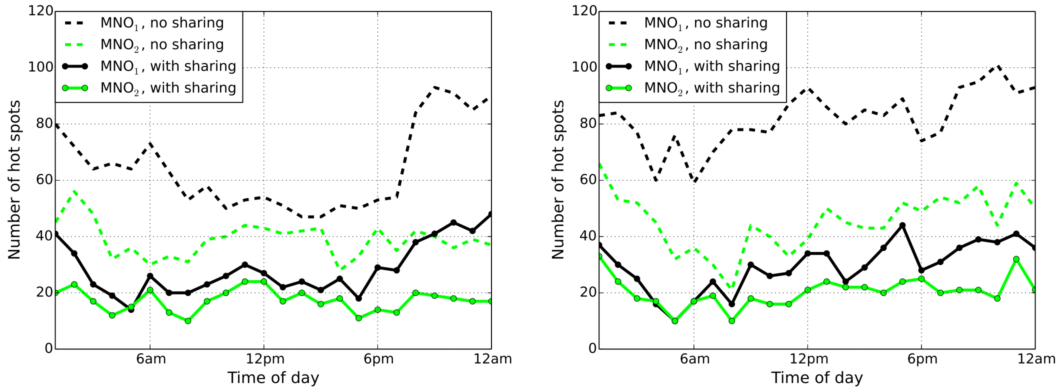


Figure 3.5: 3G data, Dublin area: number of hot spots with and without network sharing, during (a) weekdays and (b) weekends.

Broadening the focus

So far, our results have focused on the Dublin area alone. This is sensible, as Dublin is the biggest and most densely populated urban area of Ireland, and that is where network overloading issues are most likely to happen. However, next, we also present the number of hot spots nation-wide, and how network sharing can decrease them.

Fig. 3.6 shows two interesting facts. First, Dublin does not host the majority of the hot spots in Ireland. This is a bit counterintuitive, as Dublin does account for most of the traffic in Ireland.



(a)

(b)

Figure 3.6: 3G data, all of Ireland: number of hot spots with and without network sharing, during (a) weekdays and (b) weekends.

Recall, however, that the metric defined earlier in the section is *local*; it follows that hot spots in rural areas of Ireland can be, so to speak, *colder* than ordinary sectors in Dublin. Notice that, whatever their *temperature*, hot spots always represent a problem for the network.

The second interesting aspect that we can observe is that the effectiveness of network sharing in reducing the number of hot spots in rural areas is remarkable. Comparing the solid lines in Fig. 3.6 and Fig. 3.5, we can conclude that most of the rural hot spots disappear when network sharing is enabled. This is consistent with what we would expect: hot spots are fairly uncommon in rural areas, and overlapping hot spots even more so.

Table 3.2 confirms these data. Enabling network sharing removes virtually all the hotspots in rural areas, and many of the ones in Dublin. Even in the most challenging setting, i.e., weekdays in Dublin, at least one third of the hot spots can be removed.

		Operator	Ireland	Urban	Rural
Deployment density [sectors/km^2]		MNO$_1$	0.080	4.488	0.040
		MNO$_2$	0.095	5.615	0.042
Space correlation [Moran's Index]	we	MNO$_1$	0.10	0.08	0.11
		MNO$_2$	0.13	0.11	0.25
	wd	MNO$_1$	0.07	0.08	0.10
		MNO$_2$	0.04	0.04	0.11
hot spot reduction	we	MNO$_1$	-55%	-38%	-93%
		MNO$_2$	-55%	-35%	-50%
	wd	MNO$_1$	-64%	-46%	-96%
		MNO$_2$	-54%	-44%	-93%

Table 3.2: Deployment density, spatial correlation and reduction in the number of hot spots, for the whole of Ireland, urban areas (Dublin) and rural areas. Values are differentiated for weekdays (wd) and weekends (we).

3.2 Modelling Framework for Mobile Network Planning

Cellular networks are all about evaluating whether coverage and capacity constraints are satisfied. In their current form, our datasets lack several important pieces of information that are crucial to assess the performance of a network. To overcome this limitation, in this section we propose a methodology to create a low-complexity modelling framework that contains all the parameters necessary to assess coverage and capacity performance of a network. The resulting modelling framework can be used to study planning of any cellular network. We also extend its use to analyze infrastructure sharing among operators. The whole discussion is based on the Irish case, but the methodology is still valid for any other settings, provided that the right data are available.

3.2.1 Raw Data

In this section, we present the raw data that can be used to produce a real-world, large-scale cellular networking dataset, consisting of (i) cellular infrastructure deployment (ii) cellular data demand and (iii) census information. While cellular infrastructure and cellular data demand have been described in Sec. 3.1.1, below we describe the census data.

Census information. The Irish Central Statistics Office releases periodically a set of demographic and socio-economic data.[2] They are publicly available and consist of a *shapefile*, dividing the surface of the Republic of Ireland into polygons, and a *database* file, containing for each polygon, the following information:

- population, number and size of households;
- job category, income distribution at different aggregation level (i.e. individual and household level);
- age, ethnicity, language distribution at different aggregation level (i.e. individual and household level);
- classification of the area as urban, suburban, or rural.[3]

These files are available at different resolutions; however, we select the lowest possible, which comprises between 50 and 200 dwellings. They are designed in such a way that they are still in line with data protection laws.

[2] `http://www.cso.ie/en/census/census2011_boundaryfiles/`

[3] In some cases this information is explicit. In some other cases it can be inferred by looking at the density of the population per km^2 as it is done in [87].

Interesting themselves, these data become precious when correlated with network topology and demand. As an example, we could study whether a higher data demand is associated to young people (eager consumers of multimedia content, one would expect) or to wealthy areas, owing to a higher penetration of costly, high-end, high-resolution devices.

3.2.2 Creating the Adjacency Matrix

Planning a network essentially means making sure its *infrastructure* is able to serve the *demand* of its *users*. To combine the raw data described in Sec. 3.1.1 and Sec. 3.2.1 into a flexible and easy to manage description of these three elements, we design the three steps summarized in Fig. 3.7.

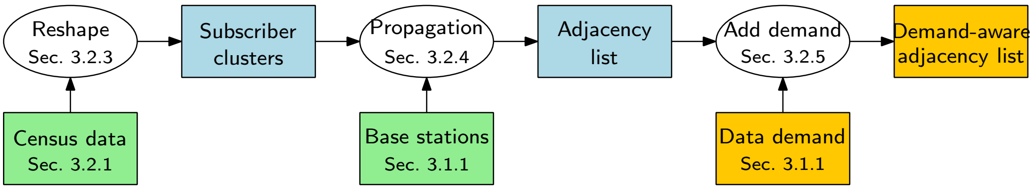


Figure 3.7: From raw data to our dataset. Boxes represent datasets; ovals correspond to models and algorithms. We begin from census data, processing them as described in Sec. 3.2.3 to obtain a list of subscriber clusters. Combining them with the location of base stations we obtain an adjacency list, as detailed in Sec. 3.2.4. Finally, we enhance the adjacency list by adding demand information, as shown in Sec. 3.2.5. Green boxes correspond to publicly available information; blue ones to information we offer for download; orange ones to information we cannot directly disclose.

We begin from users, abstracting their location through what we call *subscriber clusters*, as described in Sec. 3.2.3. Next, we turn our attention to the infrastructure, and in Sec. 3.2.4 we assess the spectral efficiency that can be achieved between each element thereof (e.g., each base station) and each subscriber cluster. Finally, we assign to each subscriber cluster its demand, using operator-provided information as detailed in Sec. 3.2.5.

Our final dataset resembles an *adjacency matrix*: for each base station and each subscriber cluster, we know:

- the position of both, and therefore the distance between them;
- the demand of the subscriber cluster;
- the capacity with which the base station can serve it.

3.2.3 From users to subscriber clusters

The format of the census information described in Sec. 3.2.1 poses two main challenges. To begin with, its resolution is too coarse. Furthermore, polygons are complex and computationally intensive to manipulate. To cope with these issues, we group users into *subscriber clusters*. Each subscriber

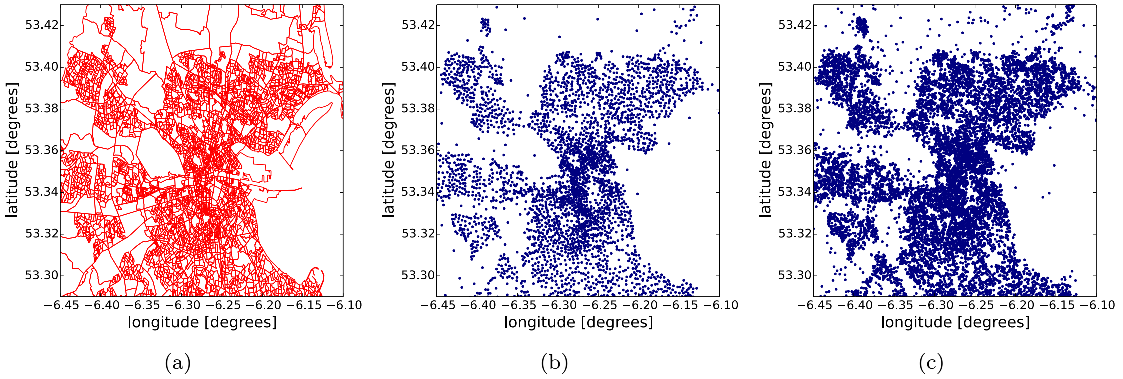


Figure 3.8: (a) The census polygons in a urban area (Dublin). Reshaping the polygons in Dublin city area using (b) the centroids and (c) setting `max_population` $= 100$ and `max_area` $= 1$ km^2.

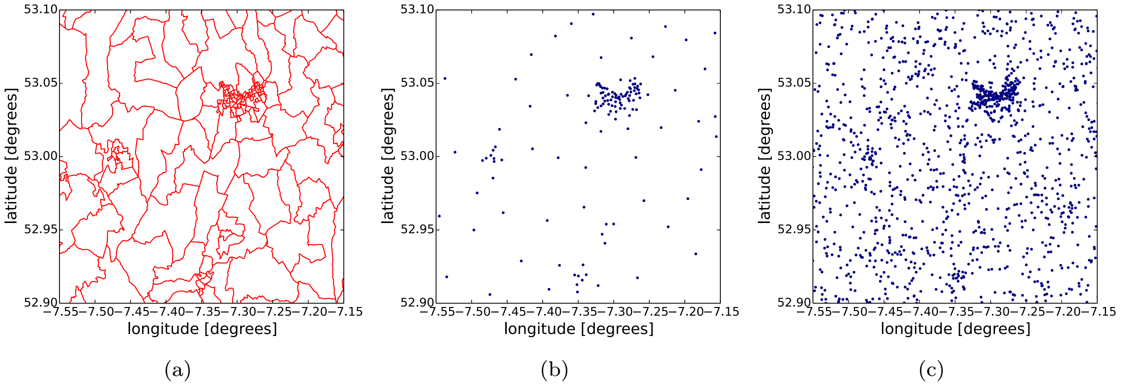


Figure 3.9: (a) The census polygons in a rural area (Laois County). Reshaping the polygons in the same rural area using (b) the centroids, and (c) setting `max_population` $= 100$ and `max_area` $= 1$ km^2.

cluster has a position in space, and represents a set of users that can be seen as co-located. More specifically, as shown in Fig. 3.8 and Fig. 3.9:

- we decide the maximum number of users and the maximum area each cluster can represent;
- for each polygon, we compute the number of clusters to place therein;
- we place the clusters randomly within the area of the polygon.

This solution has two main advantages. First, the number of clusters, the number of users and the area they represent are fully customizable and do not depend on the number and shape of the original polygons. Furthermore, the position of subscriber clusters is a point in space: computing aspects such as coverage, attenuation and spectral efficiency is simple and computationally lightweight, as noted also in [64]. It is worth stressing that the placement of demand clusters is *not* distributed according to a Poisson point process. Indeed, the location and shape of the tiles is deterministic, and given by the census data we leverage on. Additionally, the number of demand clusters we place in each tile is also deterministic, as explained earlier. The only random decision is where to locate the demand clusters within the tile, for which no further information is available.

Moreover, we have to decide the right population and area limits for our subscriber clusters. As we discussed in particular with reference to Fig. 3.8 and Fig. 3.9, lower limits mean more subscriber clusters, in both densely and sparsely populated areas. Both are important; indeed, evaluating a network planning strategy often means checking that it is able to serve all the demand from urban areas, without creating coverage problems in rural ones. However, too many subscriber clusters mean more complex simulations and longer computation times.

The total number of subscriber clusters to place in a polygon is defined as:

$$n(q) = \left\lceil \max\left(\frac{\pi(q)}{\texttt{max\_population}}, \frac{\alpha(q)}{\texttt{max\_area}}\right)\right\rceil + 1, \quad \forall q \in \mathcal{Q} \tag{3.1}$$

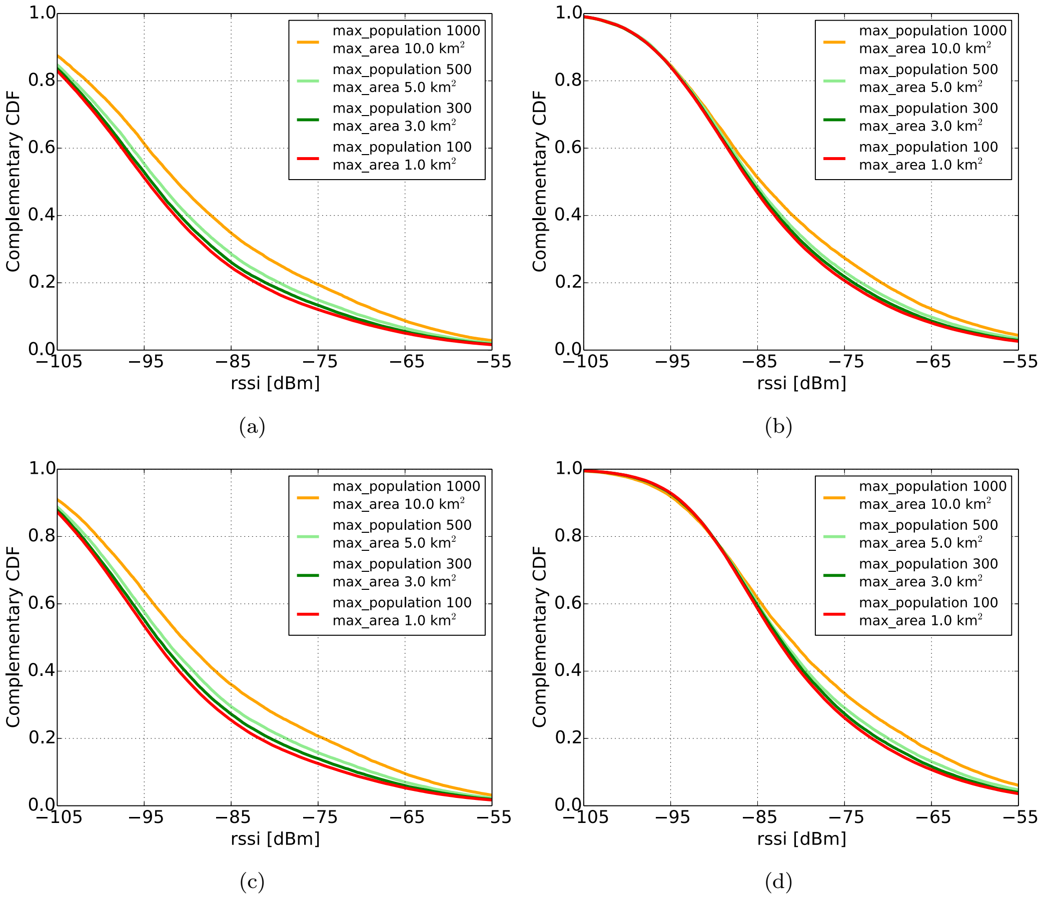
where max_population and max_area are the maximum number of people and km^2 each subscriber cluster can represent and $\pi(q)$ and $\alpha(q)$ are the actual population and area of polygon q respectively. $n(q)$ is the number of subscriber clusters to be placed in the polygon q ($\mathcal{Q}$ is the set of all polygons); these clusters are then uniformly randomly distributed within the boundaries of the polygon.

The choice of the limits max_population and max_area clearly has an impact on the accuracy of the modelling framework. Intuitively a more fine-grained representation of subscriber clusters is always more desirable but it comes at the cost of a higher (computational) complexity. Obviously there is a tradeoff to consider between the density of subscriber clusters within each polygon and the resulting accuracy.

In Fig. 3.8 we show subscriber cluster placement in an urban area. Polygons in the city centre are very heterogeneous, reflecting the different densities of the population. In this case the sampling density is dictated by the max_population input parameter and it will have an impact in assessing the capacity of the network. This is important considering that operators' main concern in urban areas is capacity shortage.

In Fig. 3.9 we show subscriber cluster placement in a rural area. Even in rural areas the polygons are not homogeneous. The snapshot in Fig. 3.9 clearly indicates that the sampling density is dictated by the max_area input parameter. In this case a more fine-grained sampling of the territory would allow a better study of the coverage, which of course is also important for operators given that coverage is their main concern in rural areas.

Fig. 3.10 and Fig. 3.11 show how the sampling density impacts the evaluation of the RSSI experienced in network. First, we evaluate when polygons are represented with a very high max_population and max_area. As the constraints are made stricter, increasing the number of points, we can notice how up to a certain level of resolution the curves collapse on each other. This is an important result



(a)

(b)

(c)

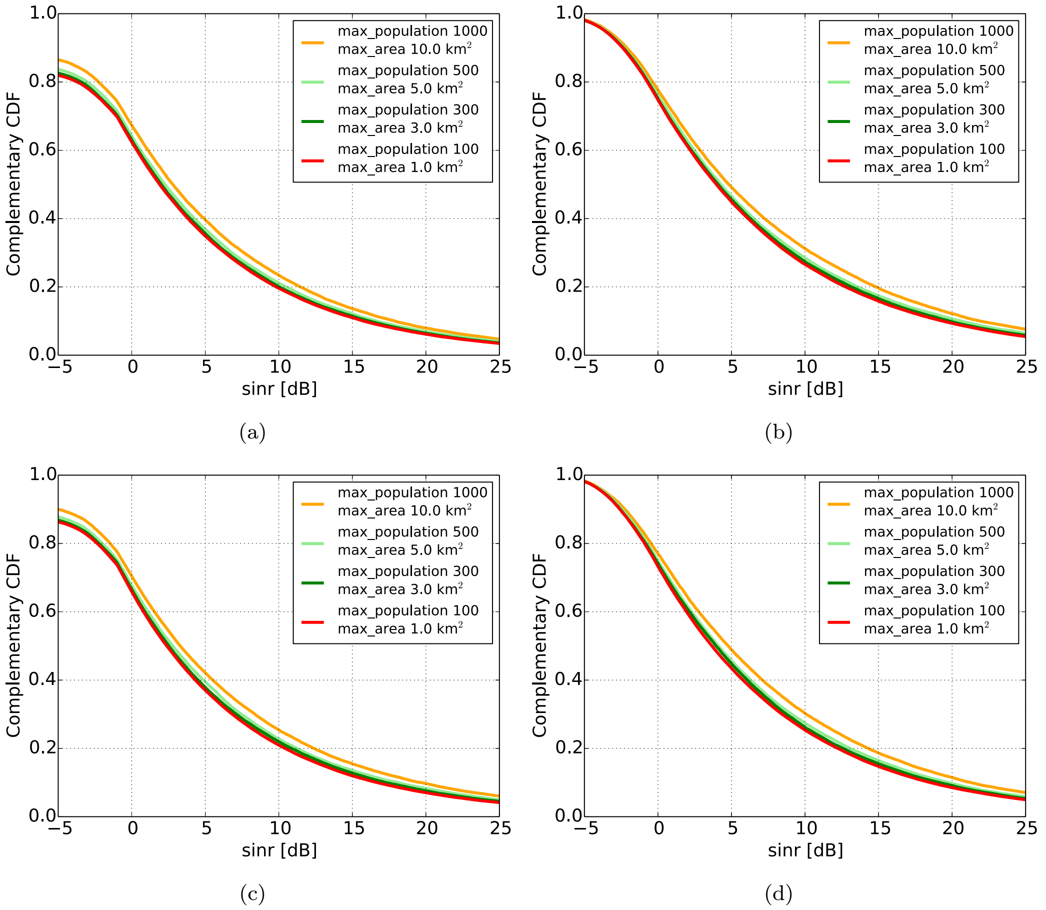
(d)

Figure 3.10: Complementary CDF of the RSSI at different sampling densities. (a) and (b) refer to MNO_1 while (c) and (d) to MNO_2. Moreover (a) and (c) refer to the 3G case while (b) and (d) to the GSM.

since it suggests that in some cases the sampling strategy can also take into account the resulting complexity of the scenario without having a substantial impact on the evaluation of the network. In our case, a population limit of 300 people and an area limit of 3 square kilometres is arguably a good compromise between accuracy and computational complexity.

3.2.4 Propagation and spectral efficiency information

The *reshaping* procedure described in Sec. 3.2.3 allows us to find the position of each subscriber cluster. In addition, we have the position of each base station, as well as the additional information described in Sec. 3.1.1. Furthermore, we know which subscriber clusters correspond to urban areas and which do not. Therefore, we are now in the position to compute the spectral efficiency that each subscriber cluster can obtain from each base station. As shown next, we proceed in three steps: computing the attenuation, obtaining the SINR values, and mapping said values to actual spectral efficiency.



(a)

(b)

(c)

(d)

Figure 3.11: Complementary CDF of the SINR at different sampling densities under the assumption that all base stations deployed by each operator interferer with each other. (a) and (b) refer to MNO$_1$ while (c) and (d) to MNO$_2$. Moreover (a) and (c) refer to the 3G case while (b) and (d) to the GSM.

Attenuation. The attenuation can be computed using one of the many models existing in literature. In our case, we opt for the COST-231 Hata model [88]:

$$L = 46.3 + 33.9 \log_{10} f - 13.82 \log_{10} h_b - a(h_r) + (44.9 - 6.55 \log_{10} h_b) \log_{10} d + c_m,$$

where f is the operating frequency; h_b and h_r are the height of base station and users respectively; $a(h_r)$ and c_m are correction factors whose values change for urban, suburban, and rural areas. Notice how this propagation model exploits the information we have about the power and frequency of base stations, properly rendering the heterogeneous nature of modern cellular networks. Our deployment infrastructure does not include antenna height information; we assume the standard value of 12 meters.

It is worth stressing that our choice of the propagation model and the parameters thereof is easily reversible. Indeed, our dataset includes the distance between base stations and subscriber clusters; therefore, a researcher who desires to use a different propagation model, e.g., the two-ray ground model, can simply do so, as detailed in Sec. 3.2.6.

RSSI and SINR. Our next step is to compute the received power (RSSI) by each subscriber cluster from each base station. This is simply the product of the transmission power of the base station and the attenuation between it and the subscriber cluster, computed as explained earlier.

The maximum power at which base stations transmit is not always available in cellular datasets. If needed, we can simply fall back to the standard values of 43 dBm for macro-base stations and 30 dBm for micro-base stations.[4]

The ratio between the power received by the transmitter and the power received by everyone else (plus thermal noise) is the SINR. We compute this value under the assumption that all base stations always transmit, i.e., full load and reuse factor 1. The reason for this very conservative assumption lies in the nature of our problem, i.e., network planning. A properly planned network has to operate in all conditions, even when facing an exceptionally high load – the infamous "flash crowds". Daily fluctuations in the load, the fact that load peaks are unlikely to happen at the same time in different parts of the topology and similar aspects can, and indeed should, be accounted for whilst operating a network, but cannot be relied upon when planning it.

As with the propagation model discussed above, this choice can be reversed by the users of our dataset: it includes RSSI values, so it is straight forward, as shown in Sec. 3.2.6, to compute the SINR under any alternative assumption if needed.

Spectral efficiency. Attenuation and RSSI can be computed, at the cost of some reasonable assumptions such as the ones we made above. Spectral efficiency, i.e., the amount of data a pair of network nodes can successfully transfer in a time unit per hertz, is either simulated or estimated. Simulation is the traditional approach: from the SINR we reconstruct the bit- or packet-error rate, and then establish whether the transmission of each packet succeeds or fails.

Owing to the scale and focus of cellular network planning, however, we adopt the other approach, and outright *estimate* the spectral efficiency from the SINR level. For example, in LTE case, we can rely on the model adopted by OfCom, based on the Shannon bound.[5] The spectral efficiency is 4.4 bits/Hz/s in optimal conditions, and reduces as the SINR decreases. OfCom themselves

[4] `http://stakeholders.ofcom.org.uk/binaries/consultations/award-800mhz/annexes/annex14.pdf`

[5] `http://stakeholders.ofcom.org.uk/binaries/consultations/award-800mhz/annexes/annex14.pdf`, Sec. A14.90

point out that their expression is a lower-bound for cellular performance, and actual deployments may exceed it.[6] As discussed earlier, adopting these conservative values suits our purpose and the objectives of cellular network planning.

3.2.5 Adding demand information

Our goal is to turn the demand information we described in Sec. 3.1.1 into a "demand" figure we can attach to each subscriber cluster. We proceed in four steps, as detailed next.

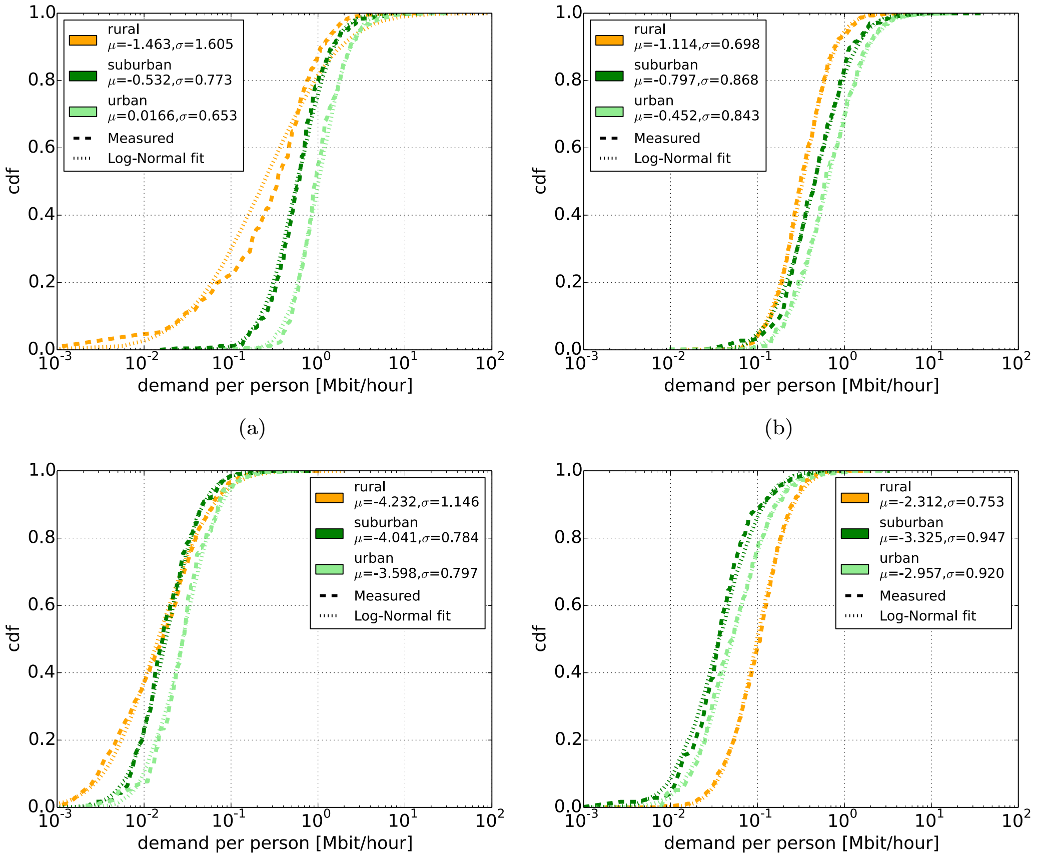
1. As traditional voice is expected to represent a decreasing percentage of the traffic that future networks will face, as a first step we restrict ourselves to the demand for mobile data.
2. The second step is aggregating the traffic over time. For each base station, we compute the data and voice load for each one-hour period, e.g., from 7PM to 8PM of November 14th, 2013.[7] This enables us to better study the time evolution of the total load.
3. The third step has to do with the nature of our problem: as discussed earlier, network planning is essentially about conservative assumptions and peak load. Therefore, for each base station, we retain the load in *its own* busiest hour, even if such hours are not the same for all base stations.
4. Fourth and last, we need to move from a *load* associated to base stations, to a *demand* associated to subscriber clusters. We do so by:
   (a) ensuring that the global load we have in our raw data corresponds to the global demand of the subscriber clusters of our dataset;
   (b) associating each subscriber cluster to the base station that provides the highest RSSI;
   (c) if a base station covers multiple subscriber clusters, its load is split proportionally to the population thereof.

Our operator data give us the *offered traffic* at each base station, at each point in time. Our key observation is that present-day and, arguably, future cellular networks will serve *all* offered traffic, by all users. In other words, each demand cluster must get at least the bitrate necessary to serve the traffic it offers during its highest-load hour, and this is what we call the *demand* the cellular network has to meet.

Now, we are able to associate to each subscriber cluster a worst-case data demand. The CDFs of *per-person* demand are summarized in Fig. 3.12. It is interesting to observe that people in urban

[6] `http://stakeholders.ofcom.org.uk/binaries/consultations/award-800mhz/annexes/annex14.pdf`, Sec. A14.95

[7] Aggregating the demand on hourly basis is a common practice in studies based on large-scale traces.



(a)

(b)

(c)

(d)

Figure 3.12: Empirical CDF of per-person 3G (a), (b) data and (c), (d) GSM demand across subscriber clusters, for urban, suburban, and rural areas for two Irish operators ((a), (c) MNO_1, (b), (d) MNO_2) alongside log-normal distribution fitting using the estimated parameters reported in the legend, i.e., location (μ) and scale (σ).

areas seem to request more data than in suburban ones. There are several possible causes for this effect, from a different penetration of high-end mobile devices to the simple availability of more capacity in densely populated areas, which encourages more traffic requests; effects like this have a major impact on network planning – and are seldom captured by smaller traces and synthetic models.

As mentioned earlier, we cannot disclose the demand figures the operators shared with us, nor can we include demand information in the dataset we make available for download. However, we do include the demand characteristics presented in Fig. 3.12 for the two operators, i.e. the CDFs of the demand for urban, suburban, and rural areas. The distributions show diversity depending on the operator and the area considered. However, most of the resulting distributions can be well approximated by the log-normal distribution whose parameters are obtained by parametric fitting with maximum likelihood estimates as shown in Fig. 3.12. This information, along with population figures from census data, can be used to reconstruct the demand at subscriber clusters level, as shown in Sec. 3.2.6 next.

3.2.6 Using our dataset

There are three ways to use our dataset: downloading it and using it as it is; customizing it to suit one's needs; or creating an entirely new dataset following our methodology.

Downloading our dataset

Our dataset is available for download form `http://bit.ly/1Fke5xV`. It consists of three files, mentioned in Fig. 3.7:

- a list of *subscriber clusters*, with their position, population, area, the Irish county they are in, and whether they represent an urban, suburban, or rural area;
- a list of *base stations*, with their position, RAT, and power class;
- an *adjacency list* containing, for each base station and subscriber cluster, the distance, attenuation, RSSI, and SINR computed as described in Sec. 3.2.4.

All datasets are in CSV format and come as compressed archives. They include a `README` file, with a detailed explanation of their format and content.

Additionally, it is also possible to download the distribution of the per-user demand in urban, suburban, and rural areas, i.e., the information shown in Fig. 3.12.

Adapting our database

As discussed in Sec. 3.2.4, our choices of propagation model and SINR-to-spectral efficiency mapping are not the only possible ones. In order to make enacting alternative choices as easy as possible, our adjacency list contains all intermediate data. As an example, users wishing to use a different mapping between SINR and spectral efficiency – e.g., because they are studying a different type of RAT, from HSDPA to 5G – can use the SINR values present in the list; users needing a different propagation model can start from the `distance` values.

Obviously changes propagate: users changing the propagation model need to recompute the RSSI, SINR and spectral efficiency values. Also notice that the format of the adjacency list is such that all operations can be performed in a *vectorized* fashion in such environments as R and MATLAB.

Adding the demand

The data set we made available also contains the demand CDFs, i.e., the data shown in Fig. 3.12. Figures are expressed in megabits, are per-user, and refer to the base station's busiest hour, as explained in Sec. 3.2.4. The demand of each subscriber cluster can be reconstructed as follows:

1. select the appropriate scenario (i.e. urban, suburban, rural);
2. extract a realization thereof, e.g., through acceptance-rejection sampling, or by approximating the spatial distribution of the traffic with a log-normal distribution using the parameters specified in Fig. 3.12;
3. multiply it by the population of the subscriber cluster.

Doing so implies the assumption that demand samples are independent, which is seldom the case. However, such a simplification is sometimes unavoidable, and is also adopted in papers proposing synthetic models [89, 90]. Exploiting the correlation with socio-demographic information can further enhance the realism of the demand profiles we obtain.

Of course, our adjacency list can be used with an altogether different demand model, e.g., one with location-specific contents.

Creating a new dataset

Information such as the one presented in Sec. 3.2.1 is increasingly easy to find. It is therefore possible to use the methodology we discussed in Sec. 3.2.2 to create an entirely new dataset, as discussed next.

National and local statistical institutes periodically release socio-economic, geographic, and demographic data. Such data are publicly available and easly accessible online. They typically come in the form of *shapefiles* (i.e., polygons in which the territory is divided) and their companion *databases*, containing polygon-specific information. Shapefiles can be easily processed with both open-source and commercial GIS softwares, and represent the input to create the *subscriber clusters*, as explained in Sec. 3.2.3.

Operator deployment data can be obtained in two ways: directly from operators themselves, or through national agencies – telecommunication regulators or health authorities.[8] They are used, along with the propagation model, to generate the adjacency list detailed in Sec. 3.2.4.

As an example, Table 3.3 summarizes where data similar to the one we used for our datasets can be found for England and Wales, Poland, and Italy. Demographic data come from national statistical

[8] This is partially an effect of mounting concerns about "electro-magnetic pollution".

institutes, while base station information are available through the national telecommunication regulators (in the case of Poland, and England and Wales) or the regional health department (in the Italian case).

Table 3.3: Demographic data and deployment data available to the public.

Country	Demographic data	Deployment data
England/ Wales	Office for National Statistics `http://bit.ly/19HkNus`	National regulator `http://bit.ly/1xnPxzL`
Poland	Geospatial portal `http://bit.ly/1qEncEg`	Office of Electronic Communications `http://bit.ly/16eLVpM`
Italy	National Statistic Institute `http://bit.ly/1fHjFJv`	Regional Environment Agency `http://bit.ly/1x9PtoO`

3.2.7 Large-scale simulator

Evaluating a network essentially means finding out whether it can comply with coverage and capacity requirements to satisfy its customers. The steps we have described to create the new dataset provide all the elements and parameters that are necessary to assess the network performance. Table 3.4 and Table 3.5 summarize the network elements and parameters, respectively, of our modelling framework. Due to the large-scale of our reference topology, we rule out traditional network simulators such as ns-2 and OMNeT++; rather, we rely on light custom simulator written in Python.

Table 3.4: Network elements of our modelling framework.

Elements	Description
$b \in \mathcal{B}$	Set of base stations
$u \in \mathcal{U}$	Set of subscriber clusters
$o \in \mathcal{O}$	Set of operators
$t \in \mathcal{T}$	Set of technologies

Table 3.5: Parameters of our modelling framework.

Parameter	Description
$d(b,u) \in \mathbf{R}^+$	(b,u) distance [km]
$\delta(b,u) \in \mathbf{R}$	(b,u) RSSI [dBm]
$\gamma(b,u) \in \{0,1\}$	(b,u) coverage range
$\gamma_I(b,u) \in \{0,1\}$	(b,u) interference range
$T(b) \in \mathcal{T}$	Technology of b
$B(b)$	Bandwidth of b
$O(b) \in \mathcal{O}$	Ownership of b
$\tau(u,o) \in \mathbf{R}^+$	Traffic demand of $o \in \mathcal{O}$ at $u \in \mathcal{U}$

In Fig. 3.13 we show the conceptual view of our simulator. Our simulator takes as input the topology information (i.e. the sets $\mathcal{U}$ and $\mathcal{B}$), the parameters estimated (i.e. $\delta(b,u)$ and $\gamma(b,u)$) and the traffic demand (i.e. $\tau(u)$), and it returns the capacity assigned to each subscriber cluster by each base station expressed as $\sigma(b,u)$. In this way, at each iteration the simulator assesses coverage and capacity in the entire topology by solving an optimization problem. The simulator will play a central role in all the problems we study in the next chapters and we will discuss its properties case by case when used.

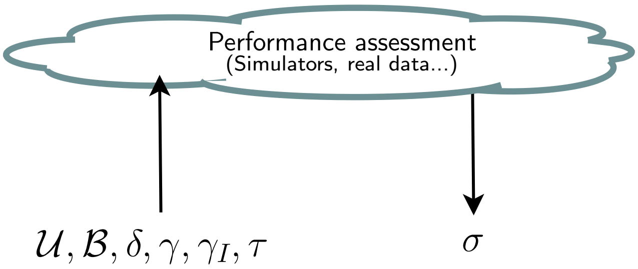

Performance assessment
(Simulators, real data...)
$\mathcal{U}, \mathcal{B}, \delta, \gamma, \gamma_I, \tau$
σ

Figure 3.13: Custom simulator implementation using real-world data.

3.3 Conclusion

In this chapter we have prepared the ground for the analysis of resource sharing in mobile networks. We started by describing the data and we have then gathered evidence that the correlation in space and time of the traffic demand between operators is low. Realizing the value of the data we have access to, we propose a modelling framework that combines operators data (i.e. topology and traffic demand) and demographic data. Finally, by using the modelling framework elements and parameters, we introduced the high level-view of the simulator we will use to assesses the performances of a large-scale mobile network.

4 Infrastructure Sharing in Mobile Networks: Consolidation and Evolution

INFRASTRUCTURE sharing is one of the simplest and most cost effective ways in which operators can extend their coverage in areas currently not served, or to increase the capacity where it is in short supply. As a consequence, we have witnessed more extreme types of sharing agreements, where two competing operators establish and manage a new consolidated network formed by their combined base stations on a national scale [91, 92].

In the first part of this chapter, we study the savings/quality trade-offs that come from network consolidation due to infrastructure sharing. We investigate the extent to which it is possible to obtain significant savings while still providing a good quality of service for the subscribers. Then, in the second part of this chapter, we argue that infrastructure sharing is a key consideration in operators' planning of the evolution of their networks. We present a framework to model this planning process while taking into account both the ability to share resources and the constraints imposed by competition regulation. We find that the ability to share infrastructure essentially moves capacity from rural, sparsely populated areas (where some of the existing infrastructures can be decommissioned) to urban ones (where most of the next-generation base stations would be deployed), with a significant increase in resource efficiency. Tight competition regulation limits to some extent the ability to share but does not entirely jeopardize those gains, while having the secondary effect of encouraging the wider deployment of next-generation technologies.

4.1 The Effect of Network Consolidation on the Savings and Quality Tradeoffs

Depending on the scenario and context, the savings attainable from network sharing coincide, in a first approximation, with the reduction in the number of base stations required to satisfy a given

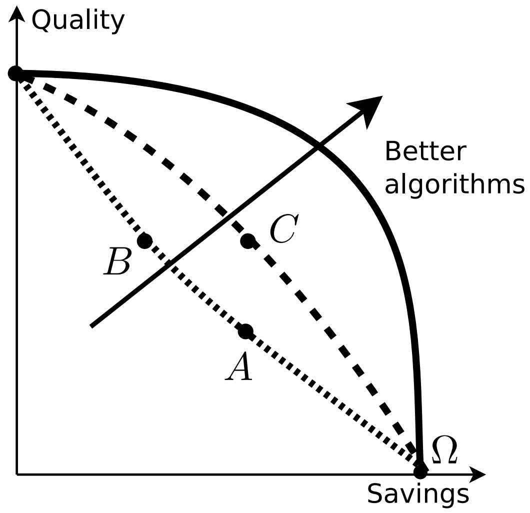


Figure 4.1: Conceptual view of network sharing. Points on the plane correspond to different trade-offs between savings and quality. Lines correspond to the sets of choices made possible by different algorithms. Point Ω corresponds to the min-cost, i.e., max-saving, feasible network configuration. Suppose we are in point A, and we want to offer our users a better quality. We can activate more base stations, thus reducing the savings we obtain (point B), or we can improve the way we choose which base stations to share, and leap to the other dashed curve. In this case, we can improve the quality without impairing the savings, moving to point C.

level of demand. Over-reliance on shared base stations may have adverse effects on the network capacity and the quality of service experienced by the users. For a mobile operator, deciding how many base stations to deploy individually, and how much to rely on shared infrastructure, is a matter of finding a compromise between savings and quality and depends upon many factors, such as contractual obligations, regulator constraints and even political reasons, some of which (e.g. competitive advantage) will be discussed later in this chapter.

However, there are other important decisions for an operator to make, for example, which existing base stations it is possible to decommission as a result of network sharing. Such a decision is of paramount importance, as it determines the price (in terms of reduced quality) the operator pays for a given level of savings due to network sharing. With reference to Fig. 4.1, each algorithm for such base station selection corresponds to a set of possible quality/savings trade-offs connected in Fig. 4.1 as a line. Improving the algorithm means pushing the line forward, making it possible to obtain *both* higher quality and savings.

A good way to improve a selection algorithm is enabling it to account for additional meaningful information. In this section, we model an operator's decision of which base stations it can potentially decommission due to the existence of shared infrastructure.

4.1.1 System model

In this section, we present our model, concentrating on the downlink.

Model elements and parameters. Our model is based on the modelling framework discussed in Sec. 3.2. In this study we use three of its elements: base stations $b \in \mathcal{B}$, operators $o \in \mathcal{O}$, and subscriber clusters $u \in \mathcal{U}$, that correspond to one or more actual users, which can be viewed as co-located, and are associated with a traffic demand $\tau(u, o)$ to be served by operator o. Furthermore, we indicate with $\tau(u) = \sum_{o\in\mathcal{O}} \tau(u, o)$ the total traffic demand at subscriber cluster u. We also introduce the ownership of base station b, expressed as $O(b) \in \mathcal{O}$.

Both base stations b and subscriber clusters u are associated with a position in space, and $d(b, u)$ indicates the distance between them. With this information we calculate the received power at each subscriber cluster u from base station b expressed by $\delta(b, u)$, which depends on the base station transmission power, the propagation model, and whether a subscriber cluster is in a rural, sub-urban, or urban area.[1]

Depending on the device sensitivity, we extract coverage information in the form of a set of binary flags $\gamma(b, u)$, expressing whether base station b can establish a useful link with subscriber cluster u. An useful link can be established when the received power at is greater the the sensitivity of the device, as expressed in Eq. (4.1):

$$\gamma(b, u) = \begin{cases} 1 & \text{if } \delta(b, u) \geq \texttt{device\_sensitivity} \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

In a similar way, in Eq. (4.2), we define the interference range $\gamma_I(b, u)$, indicating whether subscriber cluster u is in the interference range of b. By defining the interference range, $\gamma_I(b, u)$, we limit the number of potential interfering base stations by setting an interference threshold.[2]

$$\gamma_I(b, u) = \begin{cases} 1 & \text{if } \delta(b, u) \geq \texttt{interference\_threshold} \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Decision variables When mobile operators rely on a fully shared network, we model the decision of mobile operators as to which base station $b \in \mathcal{B}$ can be decommissioned because of the existence of shared infrastructure; we can express it with a binary variable $y(b) \in \{0, 1\}$, where $y(b) = 0$ indicates that base station b can be decommissioned. We also need a continuous variable $x(b, u) \in [0, 1]$,

[1] The area type of a subscriber cluster can be inferred from the population density using the demographic data described in Sec. 3.2.1.

[2] We are basically ignoring interfering base stations for which the interference power is too low.

expressing the fraction of capacity available at base station b that is assigned to subscriber cluster u. The capacity $\kappa(b,u)$ is calculated by estimating the SINR assuming a reuse factor of 1 and that all active base stations belonging to the same operator (and within the interference threshold specified in Eq. (4.2)) always interfere. In this way we set a lower bound on the capacity attainable by the network under consideration. We neglect mobility and small-scale variations such as fading and shadowing because of the nature of our problem, which is network planning [54]. In network planning, decisions are made at intervals of month or years, not milliseconds, and their effects last even longer. They are made accounting for the whole network and include, potentially, millions of users. $\kappa(b,u)$ can be expressed as follow:

$$\kappa(b,u) = B(b)log_2\left(1 + \frac{y(b)\gamma(b,u)\delta(b,u)}{\sum\limits_{\substack{b'\in\mathcal{B}:\\ b'\neq b,\\ O(b')=O(b)}} y(b')\gamma_I(b',u)\delta(b',u) + N}\right) \tag{4.3}$$

where $B(b)$ is the total bandwidth available at base station b, and N is the noise power. $\kappa(b,u)$ is the capacity base station b can offer to subscriber cluster u when b only serves u.

Constraints The first set of constraints concerns the coverage. We impose that all the subscriber clusters that, as a result of sharing agreements, can be served, can always establish a useful link with at least one base station. The coverage constraint can be expressed in Eq. (4.4).

$$\sum_{b\in\mathcal{B}} y(b)\gamma(b,u) \geq 1, \quad \forall u \in \mathcal{U}. \tag{4.4}$$

Furthermore, we cannot serve subscriber clusters other than through base stations that can establish a useful link with them:

$$x(b,u) \leq \gamma(b,u), \quad \forall b \in \mathcal{B}, u \in \mathcal{U}. \tag{4.5}$$

A second set of constraints we need to impose is that no capacity can be in decommissioned base stations, as in Eq. (4.6), and that a base stations cannot assign more capacity than it can offer, as expressed in Eq. (4.7):

$$x(b,u) \leq y(b), \quad \forall b \in \mathcal{B}, u \in \mathcal{U}. \tag{4.6}$$

$$\sum_{u \in \mathcal{U}} x(b,u) \leq 1, \quad \forall b \in \mathcal{B}. \tag{4.7}$$

Then we introduce the constraint on the traffic demand, indicating that each subscriber cluster needs to be assigned at least enough capacity to serve the traffic it carries, expressed in Eq. (4.8) as follows:

$$\tau(u) \leq \sum_{b \in \mathcal{B}} \sigma(b,u), \quad \forall u \in \mathcal{U}. \tag{4.8}$$

where $\sigma(b,u) = x(b,u)\kappa(b,u)$ represents the capacity assigned by base station b to subscriber cluster u. For simplicity, we also impose the constraint that subscriber clusters are entirely served by the active base station b with the highest received power:

$$x(b,u) \begin{cases} \geq 0 & \text{if } b = \arg \max\limits_{\substack{b \in \mathcal{B}: \\ y(b)=1}} \delta(b,u), \\ = 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

It is worth stressing that we impose constraint Eq. (4.9) only for simplicity, and that our model is able to account for more complex subscribers/base stations assignment strategies.

We will indicate with $\sigma(b) = \sum\limits_{u \in \mathcal{U}} \sigma(b,u)$ the traffic served by base station b and with $\sigma(u) = \sum\limits_{b \in \mathcal{B}} \sigma(b,u)$ the total traffic served in subscriber cluster u.

Fig. 4.2 summarizes the elements of our system model and the associated variables.

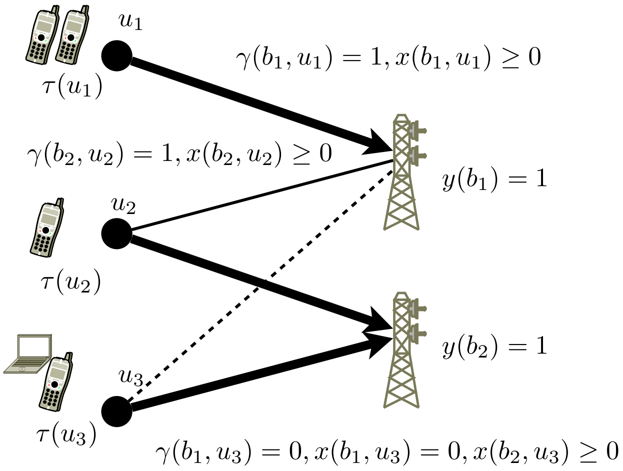

u_1
$\tau(u_1)$
$\gamma(b_1,u_1) = 1, x(b_1,u_1) \geq 0$
$\gamma(b_2,u_2) = 1, x(b_2,u_2) \geq 0$
$y(b_1) = 1$
u_2
$\tau(u_2)$
$y(b_2) = 1$
u_3
$\tau(u_3)$
$\gamma(b_1,u_3) = 0, x(b_1,u_3) = 0, x(b_2,u_3) \geq 0$

Figure 4.2: System model. Subscriber clusters $u_1 \dots u_3 \in \mathcal{U}$ on the left represent one or more users, and are associated with a traffic demand $\tau(u_1) \dots \tau(u_3)$. Parameters γ indicate coverage: subscriber cluster u_2, which is covered by both base stations, has $\gamma(b_1,u_2) = \gamma(b_2,u_2) = 1$; since base station b_1 does not cover subscriber cluster u_3 we have $\gamma(b_1,u_3) = 0$. Variables x account for service: subscriber cluster u_1 can be served by base station b_1, hence $x(b_1,u_1) \leq 1$.

4.1.2 Selection algorithms

We follow a greedy approach to determine which base stations to decommission. Intuitively, we aim at decommissioning the least useful base stations, provided that doing so does not jeopardize coverage and capacity, i.e., respects constraint Eq. (4.4) and Eq. (4.8). Alg. 1 summarizes the steps we take.

Algorithm 1 Greedy selection algorithm.

```
Require: $\mathcal{B}$,$\mathcal{U}$, $\mathcal{O}$
 1: $y(b) \leftarrow 1, \quad \forall b \in \mathcal{B}$
 2: $\mathcal{Z} \equiv \emptyset$
 3: while true do
 4:     $\mathcal{V} \leftarrow \{v \in \mathcal{B} \setminus \mathcal{Z} \colon \nexists u \in \mathcal{U} \colon \sum_{b \in \mathcal{B}} y(b)\gamma(b,u) - \gamma(v,u) = 0\}$
 5:     if $\mathcal{V} = \emptyset$ then
 6:         break
 7:     for all $o \in \mathcal{O}$ do
 8:         $\mathcal{V}_o \leftarrow \{v \in \mathcal{V} \colon O(v) = o\}$
 9:         while true do
10:             if $\mathcal{V}_o \neq \emptyset$ then
11:                 $v^\star \leftarrow \arg\min_{v \in \mathcal{V}_o} \texttt{usefulness}(v)$
12:                 $y(v^\star) \leftarrow 0$
13:                 compute $\sigma(u) \quad \forall u \in \mathcal{U}$
14:                 if Eq. (4.8) does not hold then
15:                     $y(v^\star) \leftarrow 1$
16:                     $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{v^\star\}, \mathcal{V}_o \leftarrow \mathcal{V}_o \setminus \{v^\star\}$
17:                 else
18:                     break
19:             else
20:                 break
21: return $y(b), \quad \forall b \in \mathcal{B}$
```

At the beginning of the algorithm, we assume that all the base stations are active (Line 1) and we initialize an empty set $\mathcal{Z}$ (Line 2). This set will include base stations that cannot be removed without jeopardizing the capacity constraint.

Subsequently, at each iteration we identify the set $\mathcal{V}$ of *candidate* base stations, i.e., base stations that can be decommission without compromising coverage. As we see from Line 4, for $v \in \mathcal{B} \setminus \mathcal{Z}$ to be a candidate base station, there must be no user cluster for which v is the sole active covering base station. If $\mathcal{V}$ results empty, we cannot decommission any more base stations without violating constraint Eq. (4.4), thus we return the current y-values and exit (Line 21).

Once we identify the set $\mathcal{V}$, to ensure fairness we allow mobile operators to decide the base station to decommission in turn (Line 7). Each operator has to select from the pool of its own base stations (Line 8), which one to decommission. If the current operator cannot identify any candidate base station (Line 10), we allow the next operator to select one of its base stations to decommission. In

Line 11, we identify the candidate base station $v^\star \in \mathcal{V}_o$ for which the usefulness metric is minimum. The definition of usefulness, and the elements it accounts for, have a major impact on the behavior and performance of our algorithm, as shown in the next section. Once we identify base station $v^\star$, we verify the effect of its decommission (Line 12). We need to recalculate which base stations will serve the traffic demand once served by $v^\star$ (Line 13) and whether or not the capacity constraint (Eq. (4.8)) is still satisfied. In case not, base station $v^\star$ cannot be decommissioned (Line 15), it goes to the blacklist $\mathcal{Z}$ and it is removed from the set of candidate base stations $\mathcal{V}_o$ (Line 16). Otherwise, base station $v^\star$ will be decommissioned and we allow the next operator to choose the next base station.

Usefulness metrics

As mentioned above, the usefulness metric computed in Line 11 determines which base station is decommissioned at each iteration of our algorithm, and thus the resulting network configuration and its performance. In the following, we present two usefulness metrics, differing by the amount of information they account for. Their performance is compared in Sec. 4.1.3.

Traffic-based The most natural usefulness we can attach to a base station is the volume of traffic it carries. Indeed, if a base station is used by many users to transfer large amounts of data, it is quite intuitive that to decommission it will adversely impact network performance.

```
Algorithm 2 Traffic-based usefulness metric.
Require: $v \in \mathcal{V}_o$
return $\sum_{u \in \mathcal{U}} \sigma(v, u)$
```

The expression in Alg. 2 is exactly the amount $\sigma(v)$ of traffic carried through v.

Quality-aware While traffic is arguably the most important aspect to account for in selecting which base stations to decommission, there are also other aspects to be considered. As we can see from the example in Fig. 4.3, we would sometimes prefer to decommission a slightly more loaded base station if this means avoiding serving some users with a lower signal quality. Therefore, we account for the useful received power in our metric, and decide that the usefulness of a base station depends not only on the traffic it carries, but also on how much said traffic would be affected in term of signal quality if the base station is decommissioned. This new metric is computed according to Alg. 3.

We initialize the usefulness score r to zero (Line 1). Then, for each subscriber cluster u served by base station v (Line 2), we proceed as follows. First, in Line 3, we identify the base station b'

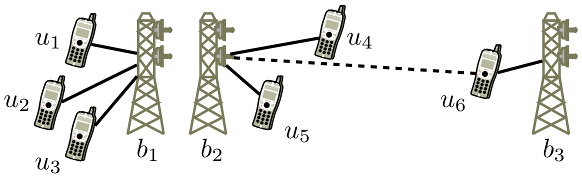


Figure 4.3: Why traffic is not the only aspect that matters. Three base stations $b_1 \dots b_3$ currently serve six subscriber clusters $u_1 \dots u_6$. All subscriber clusters have the same amount of traffic, and each station's coverage and capacity would suffice for all. Also assume that b_1 and b_2 are co-located and with the same power class. Under the traffic-based metric defined in Alg. 2, we would next decommission b_3, and u_6 would be served by b_2. While this would not disrupt coverage, it would result in a lower useful received power, hence a lower quality, for u_6. On the other hand, decommissioning b_2 would hardly trouble u_4 and u_5, and allow all users to experience a good quality service.

Algorithm 3 Quality-aware usefulness metric.

```
Require: $v \in \mathcal{V}_o$
$r \leftarrow 0$
for all $u \in \mathcal{U} \colon \sigma(v,u) > 0$ do
    $b' \leftarrow \arg \min_{b \in \mathcal{V}_o \setminus \{v\} \colon y(b)=1} \delta(b,u)$
    $r \leftarrow r + \sigma(v,u)\,[\delta(b',u) - \delta(v,u)]$
return $r$
```

that would serve u should v be decommissioned. Then, we add to v's score the product of the traffic of u currently served by v (i.e., $\sigma(v,u)$) and the *decreased* useful received power such traffic would experience, i.e., the difference between $\delta(b',u)$ and $\delta(v,u)$.

Alg. 3 is not substantially more complex than Alg. 2. The main difference is that it takes into account an additional aspect, namely, the received signal power between users and base stations. As we discussed, received signal strength is directly linked to performance, and it is our conjecture that adding the former to the picture will improve the latter.

4.1.3 Scenario and results

We rely the 3G deployment data of two Irish operators as described in Sec. 3.1.1 to populate set $\mathcal{B}$ and set $\mathcal{O}$ and the demographic data described in Sec. 3.2.1 to populate set $\mathcal{U}$ and we obtain the τ-values from the traffic demand data aggregated on a hourly basis calculated at the busiest hour for each sector as described in Sec. 3.2. In addition, we report in Table 4.1 the network parameters we use to estimate the capacity.

Table 4.1: Base stations parameters.

Base station type	Frequency	Bandwidth	Max. spectral efficiency	Max. capacity per sector	Tx power	Sectors
3G (HSDPA) macroBS	2 GHz (licensed)	5 MHz	2.5 bps/Hz/sector [93]	12.5 Mbps	40 dBm	3

`device_sensitivity` = −105 dBm.
`interference_threshold` = (`device_sensitivity` − 3) dBm.

To assess the benefits introduced by infrastructure sharing we test our algorithms against two different settings: (i) mobile operators fully share their network, and (ii) mobile operators manage their network independently.

We are concerned with three main issues: first, how infrastructure sharing affects both savings and quality; second how the usefulness metrics we discussed in Sec. 4.1.2 influence the behavior of Alg. 1; finally, how the choice of the usefulness metric impacts the overall network performance. To answer these questions, we need to be able to compare the performance of a network that is the result of shared infrastructure against the performance of networks that are operated independently. Our algorithms work unmodified, with the only difference being the sets $\mathcal{B}$ and $\mathcal{O}$.[3] The answer to the first question is provided analyzing Fig. 4.4. On the x-axis we have the savings, expressed as fraction of base stations to decommission. On the y-axis we see the average received signal strength (weighted by the traffic demand served) between subscriber clusters and base stations calculated as follow:

$$\frac{\sum\limits_{\substack{b\in\mathcal{B},\\ u\in\mathcal{U}}} \delta(b,u)\sigma(b,u)}{\sum\limits_{\substack{b\in\mathcal{B},\\ u\in\mathcal{U}}} \sigma(b,u)} \tag{4.10}$$

We immediately note that a network that is a result of the consolidation of two existing networks (green lines) provides the overall best quality as a result of the increased base station density if compared to separated networks (blue and red lines). We now look at the case where the two networks are operated independently. We note that following our procedure, both mobile operators are able to decommission some of their base stations. This result is a consequence of the fact that, since mobile operators are engaged in contractual obligations, regulations and even political reasons, mobile operators are often forced to over-dimension their networks. However, for both operators analyzed, as the number of serving base stations is reduced, the requested traffic demand is served with decreasing quality. Differently, a network that relies on shared infrastructure is able to afford up to 20% of its base stations decommissioned while keeping the same overall quality for the served traffic. This effect is due to the high redundancy present in a fully shared network that consists of infrastructure planned and deployed separately by operators to serve the same population distribution.

The second question can be answered by looking at Fig. 4.5. The x-axis reports the iteration number. Recall that at each iteration of Alg. 1 we decommission a base station; the traffic served

[3] In the case of networks operated independently, the set $\mathcal{B}$ is reduced to the 3G infrastructure that belongs to one operator and, consequently, $|\mathcal{O}| = 1$. Consequently then, the capacity constraints for the network must be adjusted to satisfy the traffic demand of the considered operator.

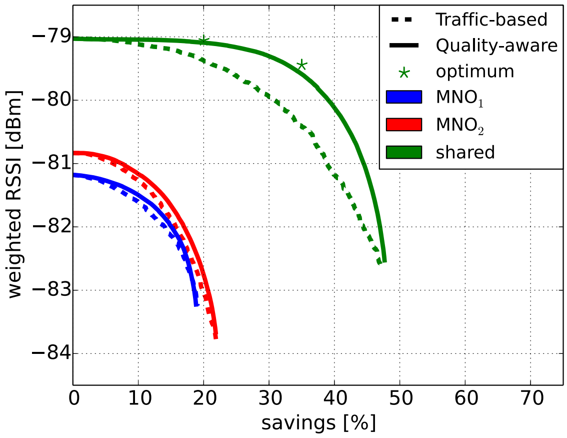


Figure 4.4: Trade-offs between savings and average (weighted) received power between users and base stations made possible by using the traffic-based and quality-aware usefulness metrics.

by said base station and its average useful received power from the subscriber clusters it serves is reported on the y-axis.

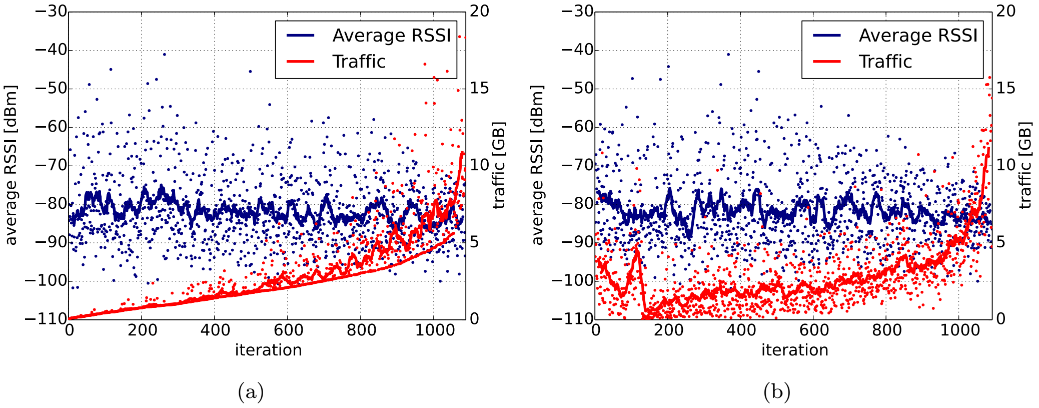


Figure 4.5: Shared case. Traffic and average RSSI from covered subscriber clusters of the base station decommissioned at each iteration under the (a) traffic-based and (b) quality-aware usefulness metrics. Dots correspond to individual iterations; lines show the moving average. The case where the two operators act independently can be found in Appendix A.

The traffic-based metric in Fig. 4.5(a) shows that the traffic demand is the only driving force of the algorithm, thus the traffic that was served by the base stations that have been decommissioned to date tends to increase monotonically. When the quality-aware metric is used, as in Fig. 4.5(b), the change is small: traffic still tends to increase, but now the decisions now also take into account the signal quality. This shows that even the quality-aware metric we present in Alg. 3 is mainly driven by the traffic, but it changes the behavior of Alg. 1 reconfiguring the network accounting for both the traffic of the base stations being decommissioned and their useful received power. Fig. 4.5 refers only to the shared infrastructure case, but the same conclusions can be drawn by looking at the networks operated independently (see Appendix A).

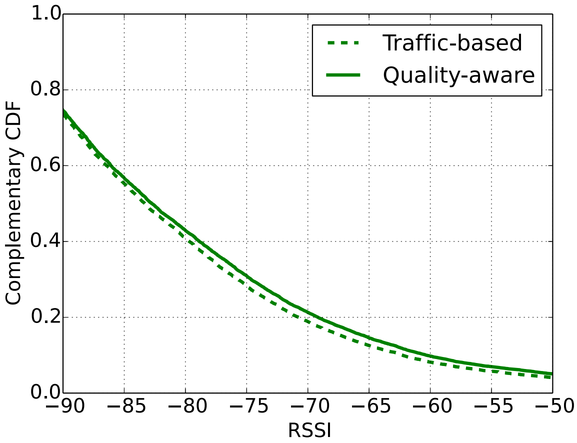


Figure 4.6: Percentage of the traffic served at a given RSSI for the shared network under the traffic-based and quality-aware metrics. The savings are set to 35%. The case where the two operators act independently can be found in the Appendix A

The third issue we posit regards the impact of the usefulness metric on the overall network performance. A first answer comes from Fig. 4.4 again. As discussed earlier, to a lower (weighted) received signal strength corresponds worse quality. The quality-aware metric is consistently associated with better performance; however, in the no-sharing cases (blue and red lines) the distance between the traffic-based and quality-aware curves is very small, while in the sharing (green lines) case the same distance is more evident. This is interesting because it suggests that operators deploy their network mainly according to their requested capacity and, as such, decommissioning base stations immediately results in an overall decreased quality for the traffic to serve. In the case of a shared network the quality-aware metric is more effective. In Fig. 4.4 we show how close the greedy approach performs compared to the optimal for two specific levels of savings, i.e., 20%, and 35%.[4] While for 20% savings the greedy algorithms performs remarkably well, as the amount of savings increases, i.e., 35%, it slightly drifts away from the optimum, likely due to the fairness criteria in place in the algorithm.

In Fig. 4.6, we look at the percentage of traffic demand being served with a certain received signal strength. We only show the shared case since similar conclusions can be derived from the no-sharing cases. We look at a specific level of saving (i.e., 35%). As expected, the quality-aware metric constantly outperforms other metric, carrying more traffic with a better quality overall.

It is also interesting to observe the similarities between Fig. 4.4 and Fig. 4.1 presented in Sec. 4.1. Switching from the traffic-based usefulness metric to the quality-aware, one has exactly the effect we labelled as "better algorithms" in that figure, i.e., we are now able to obtain better trade-offs between

[4] Due to the combinatorial nature of our problem, we only report two characteristic points calculated with an off-line algorithm.

quality and savings. In other words, whatever the level of savings we need, the quality-aware metric is an effective way of obtaining it.

One last interesting observation concerns the terminal point of the curves on Fig. 4.4, where we have the maximum savings. Unlike the idealized point Ω in Fig. 4.1, terminal points do not coincide: they correspond to different levels of savings and quality. This is an artifact of Alg. 1, which follows a greedy approach and implements a fairness criteria. As such, Alg. 1 does not find the exact optimal minimum-cost network configuration that covers all subscriber clusters, as we have also shown in Fig. 4.4.

4.2 Planning the Evolution of Cellular Networks

In this section we turn our attention to the challenges that lie in *evolving* the existing infrastructure to cope with the foreseen explosion in the demand for capacity [94]. Evolution will mean different things at different locations: some parts of the infrastructure will be replaced with new-generation equipment, e.g., LTE and its successors; others will be upgraded for the purpose of enhancing capacity, e.g., by increasing sectorization; finally, underutilized base stations will be decommissioned, possibly permanently, as part of the network consolidation process.

The yellow, solid curve in Fig. 4.7 represents the network load and its familiar predicted almost-exponential growth from the current level in A to a future one in Ω [94]. Dashed lines represent possible evolutions of the network capacity: there is no question that capacity has to increase from its current level at point B to Ω, so as to serve all the demand; in this section we investigate *how* and *when* the required changes to the network shall be performed, so as to efficiently match available capacity to demand, minimizing over-provisioning.

The reason why this matters is represented by the gray area in Fig. 4.7, corresponding to the unused network capacity. Providing capacity that nobody uses is a waste of bandwidth, resources and, ultimately, money; therefore, it is of paramount importance for operators to keep overprovisioning as low as possible. It is possible to exploit this overprovisioning through, for example, 3G onloading [95], whereby wired connectivity is augmented by cellular links. Nevertheless, it is in operators' interest to minimize costs, and therefore, deploy wireless capacity only when and where it is needed. Furthermore, Fig. 4.7 refers to the network as a whole; the relative positions of points A, B, Ω can be different in different parts of the topology. Extreme cases include some sparsely populated rural areas, where the current capacity may exceed not only the current but also the future demand, i.e., $B > \Omega$, and very dense urban areas, which may have $A \approx B$.

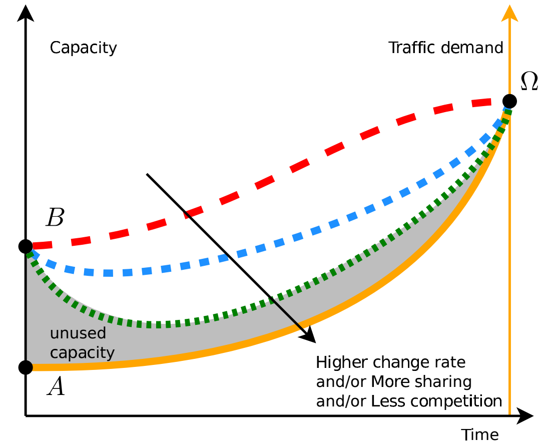


Figure 4.7: The present and future evolution of cellular infrastructures. The solid yellow line represents the traffic demand, growing from its present value in A to a much higher one in Ω. Dashed lines represent different evolutions of network capacity. Its present value (in B) is higher than the current demand A, but lower than the future demand Ω. The area between the capacity and demand curves corresponds to unused capacity (shaded in gray). Network capacity therefore has to grow from B to Ω in the long term, possibly decreasing in the short term in order to reduce the amount of unused capacity.

Ideally, operators would like their capacity to instantly fall from B to A at all locations, and would achieve this by decommissioning as many base stations as possible. Then, they would follow the demand curve all the way to Ω, by updating their infrastructure as the load increases, always keeping the unutilized capacity (i.e., the gray area in Fig. 4.7) to zero.

Such an idealized view conflicts with the reality that making any change to a network, be it deploying new base stations, updating or decommissioning existing ones, requires equipment, workpower, and funds – all resources that are scarce, and whose usage must be carefully planned. The number of such changes operators can perform in a given time is typically limited, and such a limit directly impacts the speed at which network capacity can go down or up, hence the gray area in Fig. 4.7.

Our first goal is then to study the efficient evolution of the cellular infrastructure in light of the limited budget of possible network changes. The input to our problem consists of the current set of base stations, the future demand (projected, possibly based on real-world measurements and topologies) and a limited *change rate* at which an operator can deploy, update, replace or decommission base stations. This *change rate* reflects the operators' limitations in terms of how they are able to reshape their own network. The output we seek is a list of the changes to perform to the network, and the time at which to enact each of them. The overall objective is to keep unused capacity, i.e., the gray area in Fig. 4.7, at a minimum.

In studying the evolution of cellular infrastructure, we account for two important real-world issues: network *sharing* and competition *regulation.* Both are widely studied in the literature, but their impact on the evolution of cellular networks has so far received relatively little attention.

Active network sharing [8, 96] refers to roaming-like agreements between mobile operators, where users of each operator are served through both networks indifferently. Each network operator retains ownership and control over its own spectrum. Active sharing is emerging as a promising way to achieve cost savings and enhanced performance; indeed, running networks in such a shared fashion makes it easier for operators to identify underutilized base stations to decommission as we have seen in the previous section, as well as to make the most out of updated, more highly performing infrastructure.

Network sharing agreements, in which operators actually behave as one, decrease the level of market competition as, intuitively, users have less choice. To offset this effect, regulators often require operators to leave some spare capacity, so as to allow new market players (typically, virtual MNOs) to enter the market, as imposed by the European Commission in Austria where Three Hutchison acquired the Austrian branch of Orange [92], or, more recently, when O2 and Three merged their Irish branches [91].

Studying how sharing and competition regulation shape the evolution of networks is thus an important contribution of this section.

The algorithms we present are most readily separated into three phases: meeting demand, regulation compliance, and cost reduction. Each of these phases occurs in series to update the network of an operator in order to provide service to increasing demand while minimising over-provisioning.

In our view, each individual phase conforms to an instantiation of the cognition loop. Specifically, each phase of the operation consists of observation, decision, and action steps. During observation the current situation is assessed in terms of the current network, the current demand, the already planned updates, and the expected demand. Decision involves the application of this situational awareness to some optimization. Actions take the form of changes to the schedule of network updates.

The collection of all the phases together provides a more nuanced form of cognition. This cognition uses understanding of current network infrastructure to plan future deployments based on the input of expected demand and regulatory policy. As a unit the three phases of our algorithms periodically receive an observation of projected demand, whereupon a plan for network updates is constructed. This plan is then used to decide which base station updates should be applied to the current infrastructure and the action of making these adjustments is taken.

As a result, in this section we propose a framework to study planning decisions for mobile network operators while taking into account how different aspects such as the ability to share network resources, and the constraints imposed by competition regulation impact the overall process.

4.2.1 System model

Our system model revolves around two fundamental elements: *base stations* and *subscriber clusters.*

Model elements Base stations $b \in \mathcal{B}$ are elements of the infrastructure with a certain position, capacity, and coverage area. Subscriber clusters $u \in \mathcal{U}$ can correspond to one or more actual users, which can be viewed as co-located. They have a known position and traffic demand. We also have operators $o \in \mathcal{O}$, and time periods $k \in \mathcal{K} = \{1 \ldots K\}$.

Base station type Base stations are not all equal. They can differ in technology, e.g., GSM, 3G, or LTE; furthermore, even base stations with the same technology differ in such aspects as frequency of operation and sectorization. Indeed, upgrading a cellular network essentially means changing their type, e.g., from 3G to LTE. Even decommissioning a base station can be seen as changing its type to “off”.

In our model, possible base station types are collected in set $\mathcal{T}$, and every base station $b \in \mathcal{B}$ has a type $T(b) \in \mathcal{T}$. Decommissioned base stations have the special type $t_{\emptyset}$. The type of a base station determines its coverage and performance, as shown next, as well as its associated cost.

Coverage and demand Our coverage information comes in the form of binary flags $\gamma(b, u, t) \in \{0, 1\}$, expressing whether base station b covers subscriber cluster u if b is of type t, i.e., if $T(b) = t$. Notice how these values do not depend upon time. We indicate with $\delta(b, u) \in \mathbb{R}$ the received power (RSSI) from base station b at subscriber cluster u.

Requested and served traffic For each subscriber cluster u, operator o and time period k, we know the traffic demand $\tau(u, k, o)$ from users of operator o in cluster u at period k. To streamline the notation, we will often write $\tau(u, k) = \sum_{o \in \mathcal{O}} \tau(u, k, o)$, indicating the combined traffic demand of the multiple operators expressed in Mbit.

We also indicate with $\sigma(b, u, k, o, t)$ the traffic demand that can be met by base station b, of type t, when serving users of operator o in cluster u at period k. Similarly to τ, we will often drop indices to streamline the notation, and write, e.g., $\sigma(u, k) = \sum_{o \in \mathcal{O}} \sum_{b \in \mathcal{B}} \sum_{t \in \mathcal{T}} \sigma(b, u, k, o, t)$. In the case of a sharing agreement, the combined traffic demand $\tau(u, k)$ can be served simultaneously by base stations belonging to different operators ($\sigma(u, k)$), while if no-sharing agreements are in place, each operator only serves its traffic demand using its own base stations.

Base station	$k=1$	$k=2$	$k=3$	$k=4$
b_1		↑ $x(b_1,2,t_2)=1$		↑ $x(b_1,4,t_3)=1$
b_2	$x(b_2,1,t_2)=1$	↑		
b_3	↓		$x(b_3,3,t_\emptyset)=1$	

Figure 4.8: An example schedule, with $|\mathcal{B}|=3$ base stations and $|\mathcal{K}|=4$ time periods. The maximum change rate is $N=1$, i.e., we can make at most one change (updating or decommissioning a base station) per time period. Let us assume $\mathcal{T}=\{t_\emptyset,t_1,t_2,t_3\}$, with $t_1\ldots t_3$ having increasing capacity and the same coverage. All base stations start with type t_1. Green arrows mark the periods at which base stations need to be updated due to an increased load, i.e., because $\tau>\sigma$; some stations, such as b_1, may need more than one update. We update b_1 twice, from t_1 to t_2 and then to t_3, setting $x(b_1,2,t_2)=x(b_1,4,t_3)=1$. Base station b_2 needs an update within $k=2$; However, we cannot set $x(b_2,2,t_2)=1$, because doing so would violate constraint Eq. (4.11). This forces us to anticipate the update to $k=1$, i.e., set $x(b_2,1,t_2)=1$. Similarly, the red arrow tells us that we would be able to decommission b_3 as soon as $k=1$, but the updates we already scheduled force us to delay until $k=3$, and set $x(b_3,3,t_\emptyset)=1$.

Cost Base stations also have an *operational cost* $p(b,T(b))$. Such a cost is base station- and type-dependent, and models such aspects as maintenance, site rental, and energy consumption. The cost associated with decommissioned base stations is zero, i.e., $p(b,t_\emptyset)=0,\forall b\in\mathcal{B}$.

Network changes All our decisions concern network changes. At each time period k, we may decide to *change* the type of base station b, either to a better-performing type t_{dest}, in order to increase its capacity, or to $t_\emptyset$ to save on costs, as shown in Fig. 4.8. We track type changes through binary variables:

$$x(b,k,t_{dest})\in\{0,1\}.$$

Setting $x(b,k,t_{dest})=1$ means that, at time k, we change the type of base station $b\in\mathcal{B}$ to $t_{dest}\in\mathcal{T}$. Doing nothing, i.e., never changing b's type, is represented by having $x(b,k,t)=0,\forall k,t$.

Also notice that we can change a base station's type multiple times, i.e., it can be that $\sum_{k\in\mathcal{K},t\in\mathcal{T}}x(b,k,t)>1$. We do not explicitly forbid changing the type to $t_\emptyset$ and then to some other type, i.e., first decommissioning and then re-enabling a base station, although in practice we never observe this type of behavior in our performance evaluation.

Changing the type of a base station to $t_\emptyset$ means cost saving while reducing network capacity, specifically, going down in Fig. 4.7. On the contrary, moving to a better-performing type means being able to serve more traffic, hence going up in Fig. 4.7.

In both cases, as discussed in Sec. 4.2, setting more x-values to 1 is linked to going up or down in Fig. 4.7 with a higher slope, that is being more effective in reducing the gap between requested and provided capacity.

What limits us is the maximum change rate N, defined as the maximum number of changes we can make to our network at each time period k. The following constraint must hold:

$$\sum_{b\in\mathcal{B},t\in\mathcal{T}} x(b,k,t) \leq N, \forall k \in \mathcal{K}. \tag{4.11}$$

Having Eq. (4.11) in place means two things. First and most obviously, we can take fewer actions, for instance, decommission fewer base stations. Furthermore, we may have to delay some actions, for example, decommission a base station later than we would like to. Both make us less effective in tracking the demand, i.e., imply a larger gray area in Fig. 4.7.

Time scale It is important to understand the time scale at which our model and the algorithms described later work. We are modeling network *planning*, and we are concerned with the evolution of our network over a time span of months or years. Each time period $k \in \mathcal{K}$ may correspond to several weeks, and the decisions $x(b,k,t)$ can be mapped, for example, to equipment orders or to the schedule of infrastructure deployment teams. Decommissioning or updating a base station is substantially different from turning it on and off in order to follow daily traffic fluctuation, as envisioned in "green networking" solutions [19, 20, 97]. Indeed, as discussed later, the two solutions are orthogonal and altogether compatible.

Since networks have to be provisioned for peak loads and not average ones, the $\tau(u,k)$ values express the *worst-case* amount of traffic requested by subscriber cluster u during the whole duration of time period k – for instance, the amount of traffic (in Mbit) that users in u will need served during the busiest hour of time period k. Such values typically come from forecasts and projections; from the viewpoint of our model, they are an input.

It is also worthwhile to observe that our network must be able to operate even if all the "worst hours" of all subscriber clusters take place at the same time. In other words, while it is possible, and indeed advisable, to *operate* the network so as to take advantage of the low space and traffic correlation in traffic demand as shown in Sec. 3.1, such an effect cannot be depended upon in the *planning* phase.

Assessing network performance It is important to remark that our model does not explicitly include a representation of how the traffic demand that can be met σ depends on the other parameters and variables, e.g., our decisions x. As we see in Fig. 4.9, the σ-values are obtained through an external performance assessment block. In addition to keeping our model simple, this choice affords us a higher degree of flexibility: we can interface our model with a simulation tool, or leverage any real-world data available to us, as discussed in Sec. 4.2.4.

Competition Healthy competition within the mobile market is of constant concern to regulators,

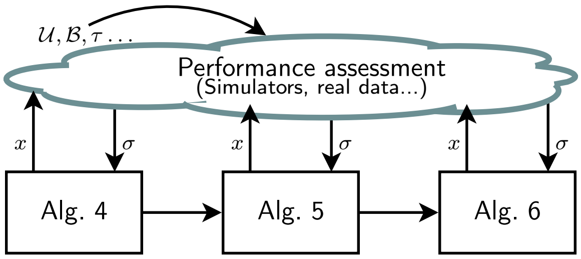


Figure 4.9: Assessing the performance within our model and solution concept. Network performance depends on topology, traffic demand and decisions; however, our model does not explicitly represent such a dependency. Our algorithms rely instead on an external block, represented by the cloud in the figure. Given as an input the network topology (base stations $\mathcal{B}$, subscriber clusters $\mathcal{U}$), demand τ, and decisions x, it returns the traffic that can be served by the network, σ, as an output. The most straightforward way of implementing such a block is network simulation; however, other approaches are possible. As discussed in Sec. 4.2.4, we implement the performance assessment block leveraging real-world traces.

as dominance on the part of large operators can lead to market abuses. A common regulatory tool to measure the level of competition and market concentration is the Herfindahl–Hirschman Index (HHI) [25, 26]. It is given by the sum of the squares of shares held by each operator in the market, and takes values between 0 (a multitude of operators with a zero-share) and 1 (a monopolist with a 100% share). The HHI can be used to assess concentration in different aspects of the market, such as, among the others, overall market share, concentration in ownership of spectrum and concentration in ownership of network infrastructure.

In our scenario, we need to define a *local* version of HHI, specific to each subscriber cluster as well as to each time period. Furthermore, we have to account not only for the operators currently in $\mathcal{O}$, but also for new operators that may enter the market if the conditions are favorable – typically mobile virtual operators (MVNOs). Our version of the HHI is thus given by:

$$H(c,k) = \left(\frac{\sigma(u,k) - \tau(u,k)}{\sigma(u,k)}\right)^2 + \sum_{o \in \mathcal{O}} \left(\frac{\tau(u,k,o)}{\sigma(u,k)}\right)^2 \tag{4.12}$$

In the denominator of Eq. (4.12) we always find the total capacity $\sigma(u,k)$ available to subscriber cluster u at time period k (recall our conventions about dropping indices). In the numerator we have the traffic demand faced by current operators in $\mathcal{O}$ in the summation, and the spare capacity, i.e., the traffic of potential new operators, in the other term.

When two operators deploy and manage their networks in a shared fashion, they behave as one from the competition viewpoint. Therefore, the set $\mathcal{O}$ shrinks, and the HHI in Eq. (4.12) increases. In our model, regulators require that the HHI not exceed a value $H_{\max}$ in at least a significant portion of the topology.

4.2.2 Problem formulation and solution

In this section, we address the problem of scheduling the network changes. Given the future demand $\tau(u,k,o)$, and the maximum change rate N, how should each operator schedule the network changes, specifically, set the x-variables? Operators have three goals:

Goal 1 – meeting the *traffic demand*:

$$\sigma(u,k,o) \geq \tau(u,k,o), \quad \forall u \in \mathcal{U}, k \in \mathcal{K}, o \in \mathcal{O}. \tag{4.13}$$

Eq. (4.13) says that for all subscriber clusters u, time periods $k \in \mathcal{K}$ and operators $o \in \mathcal{O}$, the provided capacity σ must equal (or exceed) the demand τ.

Goal 2 – complying with existing regulation:

$$\sum_{u \in \mathcal{U}} \mathbb{1}_{[H(u,k) \leq H_{\max}]} \geq \phi \cdot |\mathcal{U}|, \quad \forall k \in \mathcal{K}. \tag{4.14}$$

Eq. (4.14) imposes that, for each time period, at least a fraction ϕ of demand clusters – enough for a new operator to start building its network [16, 91] – have an HHI (as defined in Eq. (4.12)) not exceeding the limit $H_{\max}$.

Goal 3 – to minimize costs:

$$\min \sum_{b \in \mathcal{B}, k \in \mathcal{K}} p(b, T(b)). \tag{4.15}$$

An obvious way to decrease the quantity in Eq. (4.15) is setting the type of some base stations to $t_{\emptyset}$, whose associated cost is 0, that is, decommissioning them.

Multi-objective problems, where a trade-off between different goals is sought, are in general harder to formulate and to solve than single-objective problems. Thankfully, in our case, goals have a clear hierarchy: the first two goals, i.e., meeting the traffic demand, and complying with the existing regulation, must be met through as few changes to the network as possible; any remaining change can be used to pursue the third goal, that is, minimize the costs. Indeed, the first two goals can be treated as constraints, and the third one is the objective we seek to optimize.

Solution concept

Our aim is to exploit the hierarchy of the goals stated above, as well as their features, to devise a solution concept that addresses them in sequence.

We begin by defining a class of network changes that we call *capacity-preserving changes*, as follow:

Definition 1. Changing the type of a base station b from t_{orig} to t_{dest} is *capacity-preserving* if the capacity available to each subscriber cluster does not decrease. In formula:

$$\pi(b, t_{orig}, t_{dest}) = 1 \Leftrightarrow \sigma(b, u, t_{dest}) \geq \sigma(b, u, t_{orig}), \forall u \in \mathcal{U}. \tag{4.16}$$

Intuitively, capacity-preserving network changes increase the capacity available to certain subscriber clusters, without hurting others. Notice that capacity-preserving changes are also coverage-preserving. Increasing the number of sectors of a base station is a capacity-preserving change, as is replacing a GSM base station with an LTE-800 one, having the same coverage and a higher capacity. Replacing the same GSM base station with an LTE-2600 one is not capacity-preserving, as the new base station will have smaller coverage and some subscriber clusters, namely the ones covered by the old base station but not by the new one, will suffer a decrease in their available capacity. Similarly, changing any base station's type to $t_{\emptyset}$ is not capacity-preserving.

We are now in the position of proving the following useful properties:

Property 1. *Both goal 1 and goal 2 can be reached through capacity-preserving changes alone, i.e., changes that comply with Eq. (4.16).*

Proof: Goal 1 means to satisfy Eq. (4.13) for all subscriber clusters $u \in \mathcal{U}$ and time periods $k \in \mathcal{K}$. If this is not the case, then the solution is scheduling updates that increase capacity. Decreasing the capacity for some subscriber clusters, i.e., breaking Eq. (4.16), is never necessary. A similar but slightly different reasoning holds for goal 2. Increasing the σ-values, as we can see from Eq. (4.12), decreases the HHI. ■

Property 2. *If the initial configuration satisfies goal 1, then pursuing goal 2 by scheduling further capacity-preserving changes does not compromise goal 1.*

Proof: Once again, let us look at Eq. (4.13): if it holds, then all σ-values are no lower than the corresponding τ-ones, therefore, there is no way that further increasing the σ-values can change this. ■

Exploiting these properties, we propose the solution concept shown in Fig. 4.10, where objectives are addressed in sequence. Specifically, we first address goal 1, and do so by scheduling capacity-preserving network changes (Property 1 guarantees that it is sufficient). Then, we schedule further

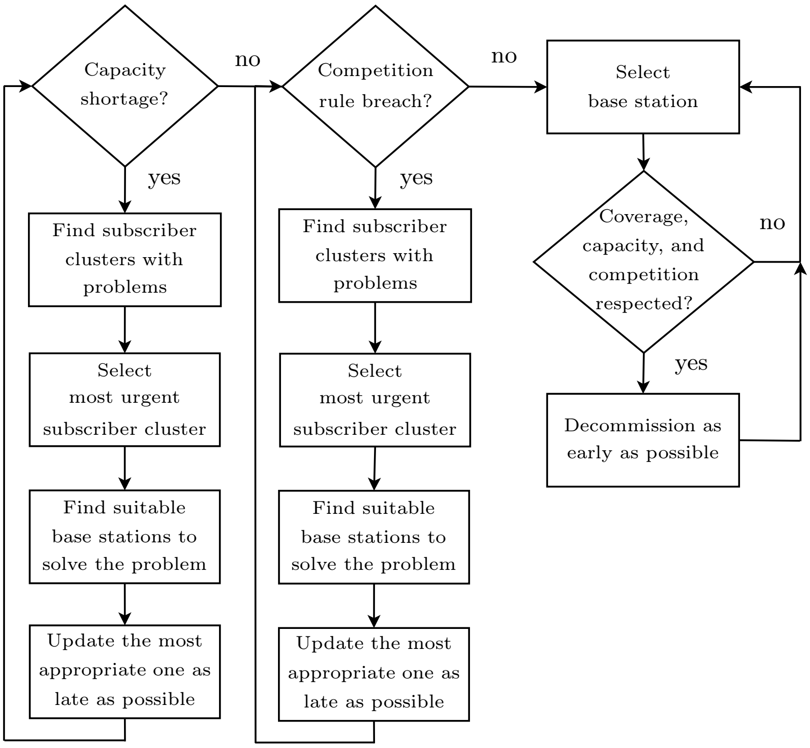

Capacity shortage?
no
yes
Find subscriber clusters with problems
Select most urgent subscriber cluster
Find suitable base stations to solve the problem
Update the most appropriate one as late as possible
Competition rule breach?
no
yes
Find subscriber clusters with problems
Select most urgent subscriber cluster
Find suitable base stations to solve the problem
Update the most appropriate one as late as possible
Select base station
Coverage, capacity, and competition respected?
no
yes
Decommission as early as possible

Figure 4.10: Our solution concept. The three phases correspond to our three objectives – meeting the traffic demand, complying with competition regulation, and reducing costs. Within each phase, we proceed in a similar way: identify the subscriber clusters that call for action, e.g., whose demand is not satisfied; identify the base stations that can be acted upon to solve the problem, i.e., whose type shall be changed, and schedule the needed action at the most appropriate time. Notice that once we are done with a phase we never come back to it, and subsequent decisions are guaranteed not to jeopardize its objective.

capacity-preserving changes in order to reach goal 2: Property 1 again guarantees that it is possible, and Property 2 makes sure that doing so will not jeopardize goal 1. Finally, we use any remaining changes we can make to the network to pursue goal 3, as long as doing so does not conflict with goals 1 and 2. Notice that in this last step we are not restricted to capacity-preserving changes, e.g., we can decommission base stations by changing their type to $t_\emptyset$.

With reference to Fig. 4.10, we can clearly see how the three goals stated above correspond to three phases in the algorithm. Within each phase, we proceed in a similar, greedy way: first we identify the problems and find the most urgent one to fix; then we identify the possible actions and find the most appropriate one; finally we schedule said action at the most appropriate time.

It is worth stressing that from our viewpoint the future demand, i.e., the τ-values, is but an input to our problem. In practical settings, such demand will not be known with precision, and this will call for appropriate action, such as, considering a safety margin. Our approach, however, remains unchanged.

Individual phases

Alg. 4 summarizes the steps we take in the first phase, where our objective is making sure that the traffic demand is met at all times and for all subscriber clusters. It works unmodified with and without network sharing: if there is no sharing, each operator will run Alg. 4 independently, feeding its own network and its own load. If operators are performing their updates in a shared fashion, then Alg. 4 will be run only once, on the joint network and the total load.

Algorithm 4 Phase 1: ensuring that traffic demand is met.

```
Require: $\mathcal{B}, \mathcal{U}, \mathcal{K}, \tau$
 1: while true do
 2:     $\sigma \leftarrow \texttt{assess}(x, \delta, \gamma, \tau)$
 3:     $\texttt{problems} \leftarrow \{(u,k) \in \mathcal{U} \times \mathcal{K} : \sigma(u,k) < \tau(u,k)\}$
 4:     if $\texttt{problems} = \emptyset$ then
 5:         break
 6:     $u^\star, k^\star \leftarrow \arg\min_{(u,k)\in\texttt{problems}} U(u,k)$
 7:     $\texttt{actions} \leftarrow \{(b,t) \in \mathcal{B} \times \mathcal{T} : \gamma(b,u^\star,t) = 1 \wedge \pi(b,T(b),t) = 1\}$
 8:     $b^\star \leftarrow \arg\max_{(b,t)\in\texttt{actions}} \delta(b,u^\star)$
 9:     $t^\star \leftarrow \arg\min_{(b,t)\in\texttt{actions}:\ \sigma(u,k,t)\geq\tau(u,k)} p(b,t)$
10:     $\widehat{k} \leftarrow \arg\max_{h=0}^{k^\star}\{h : \sum_{b\in\mathcal{B},t\in\mathcal{T}} x(b,k,t) < N\}$
11:     $x(b^\star, \widehat{k}, t^\star) \leftarrow 1$
    return $x$
```

The first thing we do is, in Line 2, to assess the performance we obtain from currently-scheduled actions, hence obtain the σ-values representing the traffic that can be served for each subscriber cluster. With reference to Fig. 4.9, calling function `assess` corresponds to entering the "performance assessment" cloud.

In Line 3, we look for *struggling* subscriber clusters, i.e., (u,k) pairs for which Eq. (4.13) does not hold. If there is no such pair (Line 4), then we are done and can move to phase 2. Otherwise, we proceed to Line 6, where we identify the $(u^\star, k^\star)$ pair that needs our attention next. The selection of the (u,k) pair to prioritize depends on the metric we decide to consider, expressed as a generic function $U(u,k)$ in our algorithm. For example, $U(u,k)$ can represent the degree of outage created by the network's problems, in which case $U(u,k) = \sigma(u,k) - \tau(u,k)$. In our case, we tackle the issue happening first, i.e., the one with the lowest k. In our case then $U(u,k) = k$.

So far, we have decided to perform a network change to tackle the capacity shortage affecting subscriber cluster $u^\star$ at time period $k^\star$. The set of base stations that we could decide to upgrade is identified in Line 7, and corresponds to the set of (b,t) pairs of base stations b such that (i) b would cover $u^\star$ if its technology were set to t, and (ii) the change would be capacity-preserving. Recall that we are relying on Property 1 and Property 2 to design the first two steps of our solution concept, and those properties only hold for capacity-preserving changes.

Among the base stations we may change, we have to identify the most appropriate one $b^\star$; in Line 8, we simply select the one that provides the highest RSSI to $u^\star$. In Line 9, we select the type $t^\star$ to update base station $b^\star$ to. We select, among the types that would restore the capacity constraint Eq. (4.13), the one with minimum cost. This also implies the mild assumption that the rate with which the traffic demand increases is never so high that we cannot restore Eq. (4.13), planning one action per time slot. Both the forecast in [94] and the fact that in our performance evaluation very few base stations are updated more than once throughout the whole simulation time, are consistent with such an assumption.

Last, we need to schedule the actual upgrade. We want to do so as late as possible, but no later than period $k^\star$. Therefore, in Line 10, we select the latest period between 0 and $k^\star$, in which we can still do something, i.e., for which we have scheduled to change the type of no more than $N-1$ base stations. Identified such a period $\widehat{k}$, we proceed with scheduling the upgrade in Line 11, by setting the appropriate x-value to 1, and move to the next iteration. In the choice of $\widehat{k}$, we can clearly see the relationship between the change rate N and our ability to keep unused capacity (i.e., the gray area in Fig. 4.7) to a minimum. Setting $\widehat{k} = k^\star$ would mean making the change when needed, hence deploying no unused capacity; being forced to have $\widehat{k} < k^\star$ means adding some network capacity that will be unused until time $k^\star$. As we clearly see from Line 10, the likelihood that we have to do so increases as N gets smaller.

It is also possible that the set in Line 10 is empty, i.e., there is no time at which we can schedule our action. This means that the network demand is growing too fast, that the change rate N is insufficient, and that network outages are unavoidable. Notice however that, as per Line 6, actions are decided in such a way to correct the earlier problems first: this means that even when outages do happen, Alg. 4 ensures they happen as late as possible.

Algorithm 5 Phase 2: enforcing competition constraints.

```
Require: $\mathcal{B},\mathcal{U},\mathcal{K},\tau$
 1: while true do
 2:     $\sigma \leftarrow \texttt{assess}(x,\delta,\gamma,\tau)$
 3:     $\texttt{problems} \leftarrow \{(u,k) \in \mathcal{U} \times \mathcal{K} : H(u,k) < H_{\max}\}$
 4:     if $|\texttt{problems}| \leq (1-\phi) \cdot |\mathcal{U}|$ then
 5:         break
 6:     $u^\star, k^\star \leftarrow \arg\min_{(u,k)\in\texttt{problems}} k$
 7:     $\texttt{actions} \leftarrow \{(b,t) \in \mathcal{B} \times \mathcal{T} : \gamma(b,u^\star,t) = 1\}$
 8:     $b^\star \leftarrow \arg\max_{(b,t)\in\texttt{actions}} \delta(b,u^\star)$
 9:     $t^\star \leftarrow \arg\max_{(b,t)\in\texttt{actions}:\ \pi(b,T(b),t)=1} \sigma(b,u^\star,t)$
10:     $\widehat{k} \leftarrow \arg\max_{h=0}^{k^\star}\{h : \sum_{b\in\mathcal{B},t\in\mathcal{T}} x(b,k,t) < N\}$
11:     $x(b^\star,\widehat{k},t^\star) \leftarrow 1$
    return $x$
```

Alg. 5 ensures that the competition constraint Eq. (4.14) is met. It has the same structure as

Alg. 4, with some differences worth highlighting. The problematic subscriber clusters, identified in Line 3, are the ones where the HHI exceeds the value $H_{\max}$. The base station type $t^\star$ to switch to is selected as the one that offers the highest capacity, so as to meet the constraint Eq. (4.14) with the smallest number of changes. Finally, the termination condition (Line 4) is triggered if at least a fraction ϕ of subscriber clusters have a sufficiently low HHI.

Algorithm 6 Phase 3: reducing costs.

```
Require: $\mathcal{B}$,$\mathcal{U}$,$\mathcal{K}$,$\tau$
 1: for all $b \in \mathcal{B}$ do
 2:     for all $t \in \mathcal{T} \setminus \{T(b)\}$ do
 3:         save$(b,t) \leftarrow \max(0, p(b,T(b)) - p(b,t))$
 4: sort save DESC
 5: for all $(b,t) \in$ save do
 6:     $\widehat{k} \leftarrow \arg\min_{h=0}^{K}\{h : \sum_{b\in\mathcal{B},t\in\mathcal{T}} x(b,k,t) < N\}$
 7:     $x(b,\widehat{k},t) \leftarrow 1$
 8:     $\sigma \leftarrow$ assess$(x,\delta,\gamma,\tau)$
 9:     if Eq. (4.13) or Eq. (4.14) does not hold then
10:         $x(b,\widehat{k},t) \leftarrow 0$
    return $x$
```

After Alg. 4 and Alg. 5, we are left with a set of capacity-preserving changes to the network that ensure that goals 1 and 2 (Eq. (4.13) and Eq. (4.14) respectively) are met. We now seek to schedule further changes, with the objective of minimizing the cost as defined in Eq. (4.15), that is, attaining goal 3. Notice that we are not relying on Property 1 and Property 2 anymore, and are thus free to schedule non-capacity-preserving changes if need be.

We proceed as shown in Alg. 6, and begin by assessing, for each base station b (Line 1) and new type t (Line 2), how much we could save by switching b's type from $T(b)$ to t (Line 3). Then we examine the potential changes, starting from the one yielding the most savings (Line 4), and simply try them out (Line 7), scheduling them at the earliest possible time period, as shown in Line 6. In Line 8 we assess the impact of the newly-scheduled change on the capacity computing the σ-values: if either Eq. (4.13) or Eq. (4.14) does not hold, i.e., if the new change impairs goal 1 or goal 2, we revert it in Line 10; otherwise, the change is confirmed.

The overall effect of Alg. 6 is scheduling further network changes, in addition to the ones decided in Alg. 4 and Alg. 5, with the purpose of reducing the operational costs as defined in Eq. (4.15). These changes are guaranteed not to jeopardize goal 1 and goal 2, thanks to the explicit check in Line 9. It is worth noting that changes are scheduled, in Line 6, as *early* as possible – conversely, the equivalent lines in Alg. 4 and Alg. 5 seek to schedule changes as late as possible. Notice that the check in Line 9 also implies that all subscriber clusters must be covered by at least one base station, i.e., no cost-saving action can be taken if that implies shrinking network coverage. As we will see in

Sec. 4.2.5, this implicit constraint has an impact on the amount of savings we can achieve.

4.2.3 Solution properties

In this section, we examine the success conditions, computational complexity and optimality of our algorithms. We prove the properties for Alg. 4, but the same holds for Alg. 5 and Alg. 6 as well, which have the same structure.

Computational complexity

Our algorithms have been designed with scalability in mind, and exhibit low complexity, linear in the number of base stations. More formally:

Property 3. *The worst-case,* combined *time complexity of all our algorithms is* $O(|\mathcal{B}|log|\mathcal{B}|)$.

Proof: At each iteration of each of our algorithms, we make exactly one decision, i.e., set one x-value to 1. Even if each base station is updated once to each possible type, the total number of decisions is still bounded by $|\mathcal{T}||\mathcal{B}|$, under the reasonable assumption that $|\mathcal{B}| \gg |\mathcal{T}|$, dominated by the sorting in Line 4 in Alg. 6, which has complexity $O(|\mathcal{B}|\log|\mathcal{B}|)$. ■

This result allows us to efficiently tackle large-scale, real-world topologies, as we see in Sec. 4.2.5. Also notice that Property 3 refers to the combined complexity of our solution concept, i.e., all the algorithms described in Sec. 4.2.2, and to the worst case; real-world cases such as the one we consider in Sec. 4.2.5 show a substantially lower complexity.

Optimality

In the following, we assess how close our algorithms perform with respect to the optimum. Our algorithms make two kinds of decisions: *scheduling*, i.e., deciding when to perform network changes, and *choosing* the changes to make. The optimality of these decisions is discussed separately.

We state and prove our properties with reference to Alg. 4, i.e., the first step in Fig. 4.10; however, since the following steps have the same structure, similar arguments hold.

Scheduling The question we examine is the following: given the times $k^{\star}$ by which changes need to be applied, how do we select the time $\widehat{k}$ at which changes are actually performed?

$\mathbf{x} = (x_k)$ represents the vector of changes we have to schedule, where x_k is the number of base stations whose type has to be changed within time period k. We begin by proving the following lemma, stating a necessary condition under which it is possible to schedule a set of changes:

Lemma 1. *The following condition is necessary for a set of updates to be schedulable:*

$$\sum_{h=1}^{k} x_h \leq kN, \forall k \in \mathcal{K}. \tag{4.17}$$

Lemma 1 can be verified by inspection of Eq. (4.17), as the total number of changes made is upper bounded by the change rate multiplied by the time in which the changes must be made. Lemma 1 says that change set that do not satisfying Eq. (4.17) is impossible to schedule. Notice that we have not proven that condition Eq. (4.17) is sufficient nor we know how to actually perform the scheduling. However, we can prove that, under certain conditions, Alg. 4 is able to schedule the changes:

Property 4. *If a set of changes satisfies Eq. (4.17), then Alg. 4 is able to schedule it.*

Proof: Scheduling a set of changes means enacting each of them at time $\widehat{k} \leq k^{\star}$ no later than its deadline. In other words, every time we reach Line 10 in Alg. 4, the set $\{h \in \mathcal{K} : \sum_{b \in \mathcal{B}, t \in \mathcal{T}} x(b,h,t) < N \wedge h \leq k^{\star}\}$ must be non-empty. In Line 6, we always select to schedule the base station with the lowest value of $k^{\star}$. This means that if at the current iteration we are scheduling a change due at time period $k^{\star}$, then all the changes we scheduled so far were due at $k^{\star}$ or earlier.

Since Eq. (4.17) holds, the number of such changes it at most $Nk^{\star} - 1$; therefore, there must be a $\widehat{k}$ between 1 and $k^{\star}$ for which fewer than N changes have been scheduled. Hence, the set is non-empty. ■

Property 4 says that Alg. 4 can schedule all sets of changes satisfying Eq. (4.17), and Lemma 1 says that all other sets of changes are impossible to schedule, regardless the algorithm. It follows that if it is possible to schedule a given set of changes, then Alg. 4 will perform it. This is important, but tells us nothing about how Alg. 4 minimizes the unused capacity, i.e., the gray area in Fig. 4.7. Specifically, if $k_b^{\star}$ is the time period at which the type of base station b needs to be changed and $\widehat{k}_b$ is the time at which the change is performed, we would like to minimize the quantity:

$$\sum_{b \in \mathcal{B}} \left(k_b^{\star} - \widehat{k}_b \right). \tag{4.18}$$

Again, Alg. 4 happens to be as effective as it gets:

Property 5. *The schedule returned by Alg. 4 minimizes the quantity in Eq. (4.18).*

Proof: We prove the property by induction.

Initialization. At the first iteration of Alg. 4, all x-values are set to 0, thus in Line 10 we have $\widehat{k} = k^\star$. The value of the quantity in Eq. (4.18) is zero, hence the schedule is optimal.

Induction step. Suppose all other changes have been scheduled optimally, i.e., they cannot be moved forward in time. Alg. 4 will try (Line 10) to schedule the current change for period $k^\star$, then $k^\star - 1$, and so on, stopping at the latest feasible time. It follows that the resulting schedule still has the lowest possible value of Eq. (4.18).

■

Choice After proving Property 4 and Property 5, we may be tempted to conclude that our approach is altogether optimal, i.e., shrinks the gray area in Fig. 4.7 to the absolute minimum. Regrettably, this is not the case: while the scheduling, i.e., deciding *when* making changes to the network, is optimal, we cannot make the same claim about the *choice* of the changes to make, e.g., the base stations whose type is to be changed.

Indeed, optimally choosing the base stations to change is an NP-hard problem. (We skip the proof, which is based on reduction from the set-covering problem.) Greedy heuristics such as the one employed in Alg. 4 are widely adopted when dealing with NP-hard problems; indeed, inapproximability results show [98] that no better solutions than the ones provided by greedy algorithms exist unless $P = NP$. In other words, the best possible polynomial-time approximation for our problem yields a solution that is no closer to the optimum (except for a constant factor) than the one of our algorithms.

Summary

From our discussion, we can conclude that our algorithms exhibit a remarkably low level of complexity, and can schedule network changes in an optimal way, that is keeping the gap between the time when a change is needed and when it is applied to the minimum.

The choice of such changes is, in general, not optimal. On the other hand, greedy approaches similar to the one we adopt are commonly used in the literature [98], and have been shown to perform remarkably well in practice.

4.2.4 Reference scenario

As done in Sec. 4.1, we study the performance of our algorithms and the factors affecting it in a large-scale, real-world scenario. In this section we briefly describe our reference topology and traffic demand, as well as the simulator we employ.

Topology and traffic demand. We leverage two demand and deployment traces, provided by two Irish operators as described in Sec. 3.1.1 collected over the whole Republic of Ireland. They include position, (approximate) coverage, and sectorization information for over 6,000 base stations, which constitute our set $\mathcal{B}$. For each base station b, the corresponding type $\mathcal{T}(b)$ (e.g., 3G, LTE) is also given. We generate the subscriber clusters using the same methodology described in Sec. 3.2 using a combination of demographic information publicly available from the Irish Central Statistics Office [99]. The set $\mathcal{U}$ corresponds to subscriber clusters generated accounting for at most (i) 300 people and (ii) at most 3 km^2.[5] Traffic demand $\tau(u, o, 1)$ at the present time slot $k = 1$ is obtained by preprocessing the traces and aggregating the traffic demand over time for each one-hour period. Since the nature of our problem is network planning problem, we retain for each base station the demand of its own busiest hour, even if such hours are not the same for all base stations. Then, by combining the preprocessed traces with the demographic data described earlier, we split the demand for each base station among all the subscriber clusters it covers according to the population each subscriber cluster covered represents. More details on this methodology can be found in Sec. 3.2. The macroscopic distribution of the traffic demand is summarized in the left plot of Fig. 4.11. Future demand is projected according to the Cisco forecast [94]. Our time horizon is $|\mathcal{K}| = 60$ time periods, with each period representing one month. The total demand for $k = K = 60$, represented on the right- hand side of Fig. 4.11, is six times the initial one.

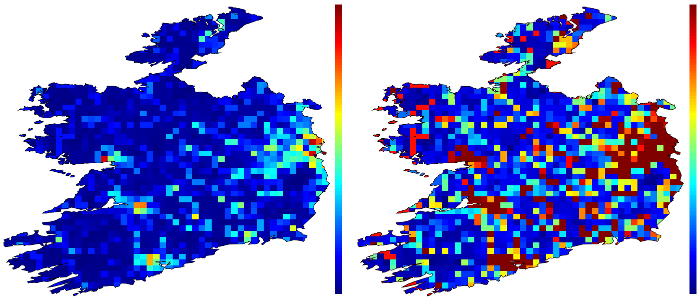


Figure 4.11: Reference scenario. Blue areas correspond to low demand, red ones to high demand. The left plot refers to the present time, i.e., $k = 1$, the right plot to $k = K = 60$ months.

Simulation and updates. The sheer scale of our reference topology rules out network simulators such as ns-2 and OMNeT++; rather, we resort to a custom simulator written in Python. We estimate the received power strength (RSSI) at each subscriber cluster using the Modified Hata model [45, Sec. 5.4.1], using the area type information to reduce inaccuracies due to overestimation and underestimation of the coverage in urban, suburban, and rural areas. The `assess` function

[5] As described in Sec. 3.2.3 300 people and 3 km^2 provide a good compromise between complexity and accuracy.

then estimates the potential achievable capacity at each time k by each subscriber cluster using the following procedure: (i) we compute the Signal to Interference Plus Noise Ratio (SINR) and the achievable spectral efficiency between any (b, u) pair and it can be calculated considering a reuse factor of 1 and that all active base stations of the same technology and operated by the same operator always interfere with each other. The maximum spectral efficiency is limited by the technology employed as specified in Table 4.2. (ii) For each technology, each subscriber cluster can be served by the base station that provides the highest RSSI. (iii) Each base station proportionally assign resources to each subscriber cluster it serves. (iv) We sum the potential capacity achievable for each technology and we obtain $\sigma(u, k)$. The minimum fraction ϕ of clusters that must experience the target competition level is set to 0.7, unless otherwise specified.

We assume that the set of base station types $\mathcal{T}$ contains the following elements:

- the *decommissioned* type $t_\emptyset$;
- a type for 3G base stations;
- three types for LTE base stations, with different sectorizations.

In Table 4.2 we present the parameters for each technology used.

Table 4.2: Simulation parameters for the different types of technologies considered.

Base station type	Frequency	Bandwidth	Max. spectral efficiency	Max. capacity per sector	Tx power	Sectors
3G (HSDPA) macroBS	2 GHz (licensed)	5 MHz	2.5 bps/Hz/sector [93]	12.5 Mbps	40 dBm	3
LTE macroBS	1.8 GHz (licensed)	20 MHz	4.4 bps/Hz/sector [100]	88 Mbps	40 dBm	$3, 6$

We assume that for each base station unitary cost since the base stations have the same transmission power.
`device_sensitivity` $= -105$ dBm.
`interference_threshold` $=$ (`device_sensitivity` $- 3$) [dBm].

The changes we can apply to the network are summarized in Fig. 4.12: we can *decommission* a 3G base station, or *create* a new LTE base station (possibly in the same location of an existing one), or *enhance* the capacity of an existing LTE base station by increasing the sectorization thereof. In the following, we will collectively refer to the last two operations as *updates*. It is worth stressing that these limitations are not inherent to our model, which is able to account for any kind of network update and to interface with any simulator (see Fig. 4.9), but merely a way to simplify (and speed up) our performance evaluation. Also notice that our model is able to account for the deployment of new base stations, which corresponds to an update from $t_\emptyset$ to any other type. Similarly, additional radio access technologies such as “small cells” would correspond to extra values in $\mathcal{T}$, and new branches in Fig. 4.12.

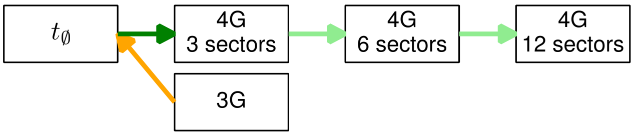


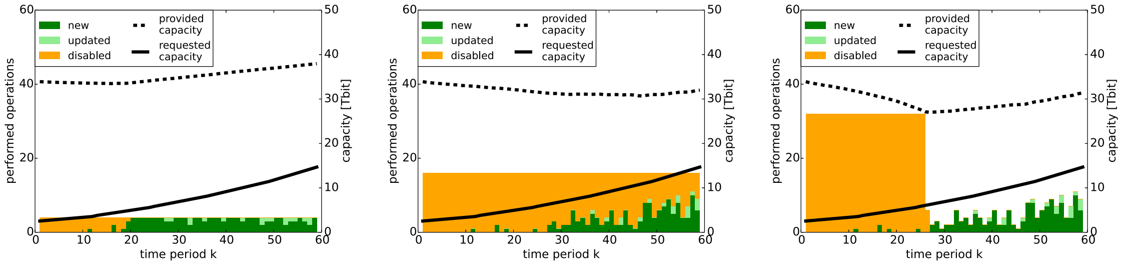
Figure 4.12: The base station types in $\mathcal{T}$ and the possible changes. 3G base stations can be decommissioned (orange arrow), i.e., have their type set to $t_\emptyset$. New LTE base stations can be created (dark green arrow), possibly in the same location as existing base stations, or have their capacity enhanced through sectorization (light green arrows).

4.2.5 Results

We begin by looking at which network changes are performed at each time period k, and how they impact network capacity. In Fig. 4.13, solid and dotted lines correspond to requested traffic τ and provided capacity σ respectively; bars represent the number of created, enhanced and decommissioned base stations. We are setting in the most favorable case: networks can be operated jointly and there is no competition constraint, i.e., $H_{\max} = 1$.

Fig. 4.13(a) represents the case for $N = 4$, the minimum possible value of N in our scenario, as given by Eq. (4.17). It is easy to see that the value of N directly maps to the maximum height of the bars. Very low values of N, as in Fig. 4.13(a), imply that most of the changes operators are able to perform are updates (i.e., create or enhance base stations), so as to meet the demand goal Eq. (4.13). As N increases, as in Fig. 4.13(b), we are able to decommission more base stations, and to push forward in time all the updates. For even larger values of N, we see that there are some time periods when we perform fewer operations than we could, i.e., $\sum_{b\in\mathcal{B},t\in\mathcal{T}} x(b,k,t) < N$, as it happens for $k > 25$ in Fig. 4.13(c). This is because we scheduled to decommission all possible base stations at earlier times, as mandated by Line 6 in Alg. 6, and schedule all needed updates later in time, as in Line 10 of Alg. 4.

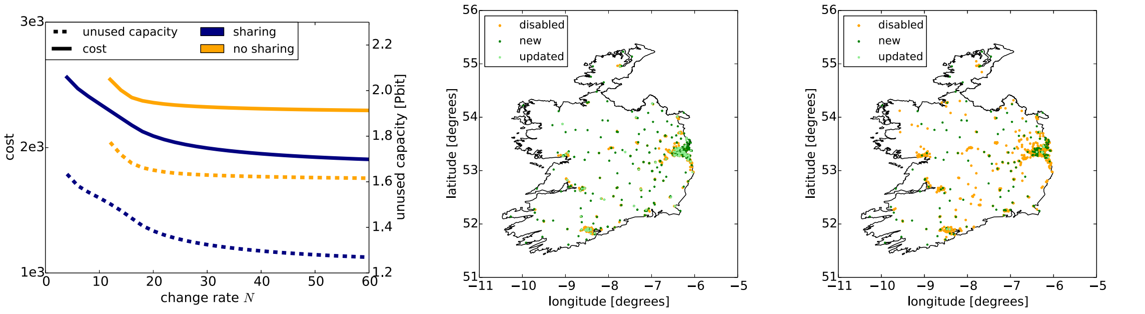
Looking at provided and requested capacity, we can notice that the provided capacity is always substantially higher than the demand. This is because we have to preserve the coverage in the entire topology, i.e., all subscriber clusters $u \in \mathcal{U}$. In sparsely populated areas, this inevitably translates into underutilized base stations that operators have no way to decommission. It is also interesting to see how N influences the evolution of provided capacity: low values of N imply that the capacity slowly increases as updates are performed (Fig. 4.13(a)). Higher values of N, as in Fig. 4.13(b), mean that we can observe the behavior we were expecting in Fig. 4.7, with network capacity first slowly decreasing due to decommissioning base stations and then leveling up due to the concurrent scheduling of updates and decommissions. As we can see from Fig. 4.13(c), further increasing N implies that network capacity decreases more swiftly (as operators can be quicker at decommissioning base stations) and increases more quickly afterwards, as most updates take place. Both effects are consistent with our intuition and expectations (Fig. 4.7): being able to perform more changes to the network means being more effective in tracking the traffic demand.



(a) (b) (c)

Figure 4.13: Changes applied to the network and requested and provided capacity for each time period k, when (a) $N = 4$ $(= N_{\min})$, (b) $N = 16$, (c) $N = 32$.

In Fig. 4.14 we look at the benefits of sharing, i.e., what savings operators can obtain by operating and updating their networks in a shared fashion. Fig. 4.14(a) is fairly clear – the benefits of sharing are very significant. The main effect of allowing sharing is that operators can save substantially more on operational costs, and thus, decommissioning more base stations. Sharing also reduces the unused capacity, which however remains quite significant, due to coverage requirements, as we already observed from Fig. 4.13.



(a) (b) (c)

Figure 4.14: (a) Unused capacity and cost savings (as defined in Eq. (4.15)) as a function of N with and without sharing; location of updated and decommissioned base stations (b) without and (c) with sharing.

The maps in Fig. 4.14(b) and Fig. 4.14(c) show *where* base stations are updated and decommissioned. Focusing on Fig. 4.14(b), which refers to the case where no sharing is allowed, we can observe that most updates are concentrated in densely populated areas (e.g., Dublin in the East), but some take place also in rural areas. On the other hand, virtually all decommissioned base stations are located in rural and suburban areas. Allowing sharing and moving to Fig. 4.14(c), we see a very different picture. In rural areas operators have to update much fewer base stations and can decommission many more; furthermore, many base stations can be decommissioned also in urban areas, e.g., in Dublin.

Put together, these results confirm the intuitive notion that network sharing directly translates into better network efficiency. Backed by our real-world demand and deployment traces, we can add that said efficiency is mostly attained by decommissioning underutilized base stations and, to a lesser extent, by pooling updated ones. Location-wise, we can say that network updates have the overall effect of *migrating* capacity from rural areas to urban ones, and sharing makes such an effect more pronounced, taking advantage of the redundancies of deployments, in particular in the rural areas.

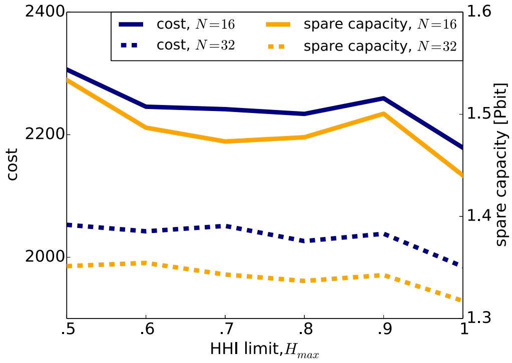
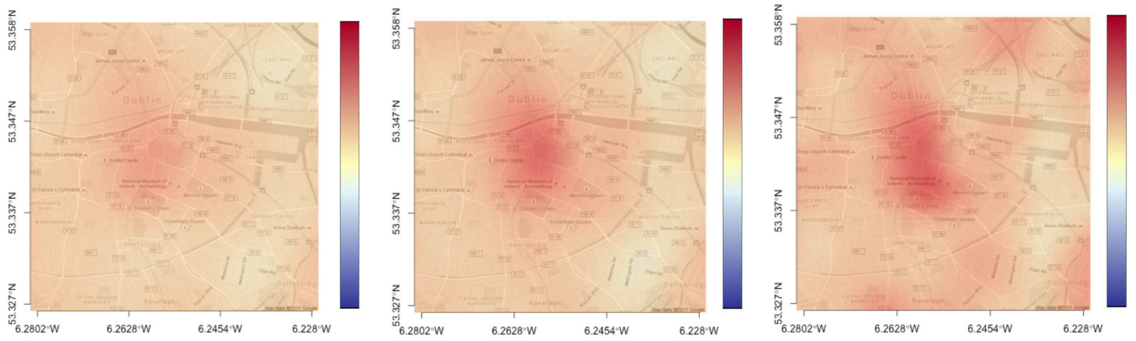


Figure 4.15: Unused capacity and cost (as defined in Eq. (4.15)) as a function of HHI index and for different values of N.

Competition regulation brings its own requirements on network planning, in particular mandating a certain level of extra capacity which could be used by a potential new virtual provider. In Fig. 4.15, we investigate the effects of competition regulation on the evolution of the network infrastructure. Recall that $H_{\max} = 1$ means that there is no regulation in place, while $H_{\max} = 0.5$ corresponds to the most stringent regulation, imposing as much as 50% idle capacity.

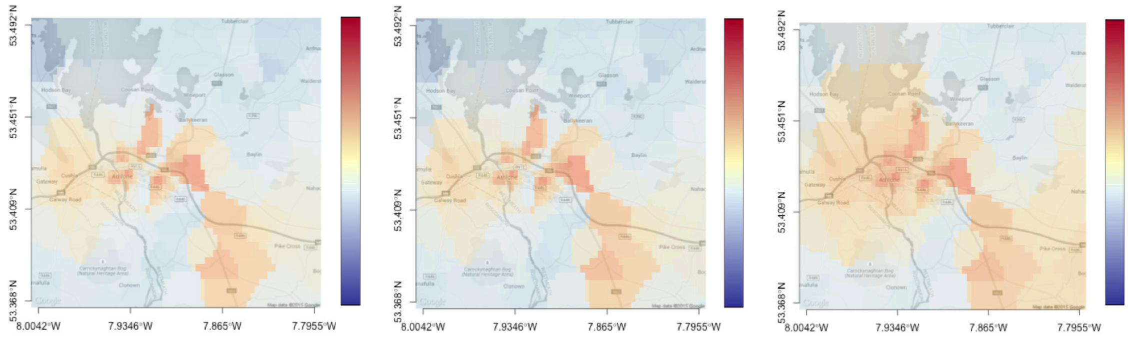
Fig. 4.15 confirms that the tighter the competition regulation, the lower the savings operators are able to achieve. As expected, moving from the most loose to the tightest level of regulation also increases the unused capacity – i.e., from the regulator's viewpoint, the capacity available to new operators. In fact, imposing a minimum competition level has a double effect: first, it forces the operators to perform some updates that otherwise would have not been necessary to meet the capacity goal in Eq. (4.13); second, it restrains the operators from decommissioning some underutilized base stations. Intuitively, since most of the decommissioning would take place in the rural areas and most of the updates in cities (Fig. 4.14(c)), the regulation has the effect that users from both sparsely and densely populated areas are able to choose between more (actual or potential) operators.

While Fig. 4.15 gives a macroscopic view of the overall capacity available, in Fig. 4.16 and Fig. 4.17 we compare how the competition affects the LTE capacity supplied in two different localized areas, i.e., a densely populated urban area in Dublin city center, and a sparsely populated rural area respectively. Fig. 4.16 and Fig. 4.17 confirm our previous findings. Without regulatory constraints, the operators still make investments to upgrade their infrastructures in densely populated areas, likely the ones with higher returns of investment (ROI), see Fig. 4.16(a) and Fig. 4.16(b), while in rural area they do not have incentives to make additional investments, see Fig. 4.17(a) and Fig. 4.17(b). On the other hand, regulatory constraints have the effect of stimulating more upgrades in both area, see Fig. 4.16(c) and Fig. 4.17(c).



(a) (b) (c)

Figure 4.16: Dense urban case, Dublin - (a) initial LTE capacity, (b) final LTE capacity when $H_{\max} = 1$, and (c) $H_{\max} = 0.5$. Values are presented normalized.



(a) (b) (c)

Figure 4.17: Rural case, Co. Westmeath - (a) initial LTE capacity, (b) final LTE capacity when $H_{\max} = 1$, and (c) $H_{\max} = 0.5$. Values are presented normalized.

Fig. 4.18 gives us further insights on the effect of competition regulation on LTE and 3G coverage. Specifically, it shows the changes in the coverage by presenting the complementary CDF of the subscriber clusters' RSSI in the original LTE and 3G deployment (i.e., $k = 1$), and at the end of the time horizon (i.e., $k = K = 60$) when there is low and high competition in place, e.g. $H_{\max} = 1.0$ and $H_{\max} = 0.5$ respectively. Fig. 4.18 shows that higher competition stimulates the deployment of newest technology in areas that otherwise the operators would not find attractive, increasing both the number of subscriber clusters with a minimum signal strength and augmenting the overall signal quality. On the other hand, since the regulator enforces a minimum level of coverage for each technology, we do not observe a decrease in the number of subscriber clusters served by the older technology, but rather a small decrease in the signal quality, mainly due to the high redundancies existing in the infrastructure at $k = 1$.

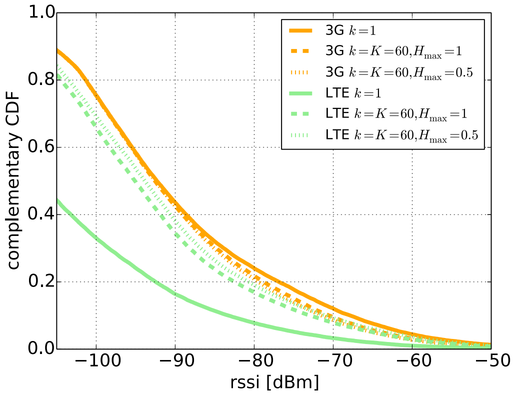


Figure 4.18: Complementary CDF of the RSSI for LTE and 3G at time $k = 0$, and at $k = K = 60$ with low competition and high competition.

4.3 Conclusion

In the first part of this chapter we considered the scenario where operators are able to obtain cost savings due to employing shared infrastructure. The trade-off between savings and quality is a primarily commercial decision; however, selecting *which* base stations to decommission is an interesting technical problem.

We assumed a greedy decision procedure, where at each iteration we decommissioned the base station deemed to be the least useful. We showed that the usefulness metric employed has a major impact on the quality of the decision being made. If operators decide to pool their infrastructures in a network sharing fashion then both cost savings and quality can improve.

In the second part of this chapter, we extended the results obtained in the first part and we focused on the *planning* and the *upgrading* of a shared network. We studied the modernization phase cellular networks will go through in the near future: mobile operators will decommission some underutilized base stations in order to save on costs, and deploy new-generation base stations in order to cope with the increasing demand. Operators can join forces and perform said changes in a shared way, so as to improve the efficiency of their networks. Operators' ability to share infrastructure may be constrained by competition regulation, and the speed with which operators can make changes to their own networks may also be limited by practical considerations. We incorporate both factors in out study of cellular network planning.

Clearly, an operator's decisions about upgrading and decommissioning infrastructure will take into account a complex mix of technical and economic factors, from the OPEX associated with different radio access technologies to possible market advantages over one's competitors. Our model aimed to capture the technical constraints on coverage and capacity that can be viewed as a first step in this decision making. As such, we use the number of base stations currently deployed as a rough proxy for the cost faced by the operator. A more sophisticated economic model is one avenue we intend to pursue in the continuation of this work.

We presented a general framework that describes network modernization scenarios, accounting for real topologies and demand information, including multiple base station technologies. This model was presented in Sec. 4.2.1. Given our model and a limited budget of changes to perform at each time period, we presented in Sec. 4.2.2 a family of algorithms able to schedule the changes in a cost effective manner, while satisfying the demand and complying with regulatory constraints. These algorithms work unmodified whether operators perform their updates individually or in a shared fashion and, as shown in Sec. 4.2.3, return quasi-optimal solutions with a very small computational complexity, dominated by their sorting stage.

We apply our algorithms in a large-scale scenario, built from real-world demand and deployment traces as described in Sec. 4.2.4. As summarized in Sec. 4.2.5, we found that network modernization essentially means moving capacity from rural, sparsely populated areas (where many base stations can be decommissioned) to urban ones (where most of the new-generation base stations are located). Allowing sharing, i.e., permitting operators to jointly update and manage their networks, greatly enhances their effectiveness. Such benefits are reduced if tight competition rules are in place, but never entirely jeopardized. Indeed, tight competition regulations have the secondary effect of stimulating operators to extend the capacity and coverage of their new-generation networks, which can therefore serve a larger fraction of their demand. As a result, our model is able to capture the tradeoff between *savings* and *the promotion of innovations* which is one of the main goals of regulators in the telecommunication industry.

5 Sensitivity Analysis on Service Driven Network Planning

THE mobile market is rapidly changing and becoming more complex. Nowadays mobile network operators' (MNOs) ability to generate revenue relies, firstly, on their subscribers and, secondly, on wholesale agreements with mobile virtual network operators (MVNOs) in a second-tier market [101]. Unfortunately this revenue model does not seem sustainable as the growing demand for capacity and data-rates forces MNOs to heavily invest in costly network infrastructure expansions and upgrades, endangering the profitability of running a mobile network. As a result, MNOs are showing interest in different business models [102].

Meanwhile, many over-the-top service providers (OTTs) have based their success on the users' perception of *limitless* traffic [103] and internet's *ubiquitous* access. Mobile capacity shortages, and subsequent service degradation, would affect OTTs' ability to generate profit. In particular, the OTTs offering bandwidth-intensive services such as HD video streaming on-demand or online gaming, which require strict quality of service (QoS), are the most exposed. Essentially, these OTTs are presented with two (non-exclusive) strategies: (i) to acquire capacity on-demand from MNOs, and (ii) to deploy their own infrastructure. Indeed we are already starting to witness similar scenarios. For example, the recently unveiled Google's Project Fi [41] offers to its subscribers both Wi-Fi, as part of Google's effort to deploy its own infrastructure, and LTE connection, as part of Google's MVNO agreement with traditional MNOs (i.e., Sprint and T-Mobile in the US). Other examples also exist and include the internet.org initiative by Facebook [104] and the Twitter deals [44].

In this model, OTTs can decide to enter into service level agreements (SLAs) with an MNO to meet the expected QoS for their services. In exchange for a fee, the MNO will reserve enough capacity to satisfy the QoS expected.[1] The OTTs would need to decide whether it is more cost

[1] As an example, the Netflix-Comcast deal in the US [105] was stipulated to increase the video streaming quality of Netflix's customers.

effective to rely on SLAs with selected MNOs or to deploy their own network infrastructure. The MNOs, in turn, would factor SLAs with OTTs into their decision of whether and how to expand their networks. In other words, we will enter the age of *service-driven network expansion*, and, more forward-looking, *service-driven networks*.

In order to study service-driven network expansion we need to assess, first, which factors are likely to influence SLAs, and second, the characteristics of the resulting networks. The former are presented on the left hand side of Fig. 5.1 and include technical and non-technical aspects. Factors considered include the technologies available (e.g. LTE, WiFi) and their costs, public policy and regulation (e.g., whether to release new bands to the public, spectrum licensing schemes), and the characteristics of the demand. The resulting network characteristics are presented on the right hand side of Fig. 5.1 and include, for example, the level of heterogeneity of the resulting network in terms of both ownership and technology, the use of licensed/unlicensed spectrum, and the emergence of virtual networks tailored to OTTs. The likely end result is a move from the current paradigm, where networks are designed, owned and controlled by MNOs, to a new one, where OTTs have a major role in the deployment of new infrastructure. Infrastructure will tend to become more heterogeneous, and integrate different equipments, some of which will operate on unlicensed spectrum (e.g., ISM bands, as in LAA-LTE [106]).

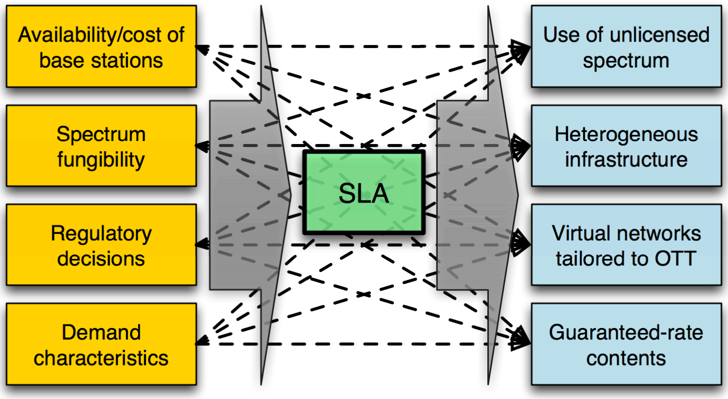

Availability/cost of base stations
Spectrum fungibility
Regulatory decisions
Demand characteristics
SLA
Use of unlicensed spectrum
Heterogeneous infrastructure
Virtual networks tailored to OTT
Guaranteed-rate contents

Figure 5.1: The features of service-driven networks (right hand side), and the factors driving them (left hand side). Dashed lines and large grey arrows represent the aspects and relationship we intend to study.

The purpose of this chapter is to investigate, qualitatively and quantitatively, the impact of the different factors, listed on the left-hand side of Fig. 5.1, on the SLAs between MNO and OTTs and on the planning decisions in service-driven networks.

The first contribution of this chapter is a synthetic model of cellular network deployment that accounts for some of the most relevant aspects of service-driven network deployment described in Fig. 5.1 and how they interact with each other. This model captures the decisions made by OTTs on whether or not to deploy their own infrastructure or to enter into an SLA with MNOs and it is presented and discussed in Sec. 5.1. Then, in Sec. 5.2, we detail our solution concept, showing how the main actors involved in the network expansion process, specifically, traditional MNOs and OTT providers, efficiently make self-interested, near-optimal decisions. Sec. 5.3 contains estimates for all the factors we account for (i.e., all the boxes on the left hand side of Fig. 5.1). Results, obtained for the real-world topology described in Sec. 5.4 and summarized in Sec. 5.5, show how these factors impact how and by whom service-driven networks will be built and operated.

5.1 System model

In this section, we present our system model, summarized in Fig. 5.2. We consider a *snapshot* of the network, taken during high-load conditions that are typically [45, Sec. 10.3.3.2] used as a reference when planning a network. The purpose of our model is to capture the network conditions in a challenging situation (as detailed in Sec. 5.3 and Sec. 5.4), so its infrastructure can be planned accordingly. In this study, we also assume that backhaul is not a limiting factor for network capacity.

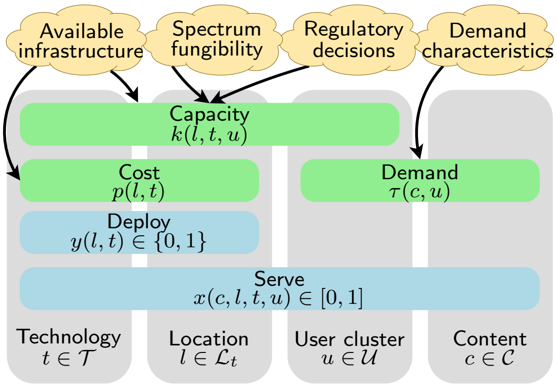

Available infrastructure
Spectrum fungibility
Regulatory decisions
Demand characteristics
Capacity $k(l,t,u)$
Cost $p(l,t)$
Demand $\tau(c,u)$
Deploy $y(l,t) \in \{0,1\}$
Serve $x(c,l,t,u) \in [0,1]$
Technology $t \in \mathcal{T}$
Location $l \in \mathcal{L}_t$
User cluster $u \in \mathcal{U}$
Content $c \in \mathcal{C}$

Figure 5.2: Our system model. Grey, vertical blocks represent the model entities. Horizontal blocks correspond to parameters (green blocks) and decision variables (blue ones). Horizontal and vertical blocks cross when the parameter/variable represented by the horizontal block is indexed by the entity represented by the vertical one, e.g., a deployment decision is taken for each technology and base station. Yellow clouds, corresponding to the boxes on the left hand side of Fig. 5.1, indicate the sources of our parameters.

The level of abstraction of our model is one of the most critical decisions we have to make. Cellular networks, especially next-generation ones, are highly complex entities, and any model trying to capture all this complexity would be exceedingly difficult to handle. As such, we employ simplified models to estimate, for example, the capacity for the different technologies, or the candidate locations of new base stations. However, it is important to stress that our goal is *not* to propose a comprehensive model of cellular networks, but rather to study the relationships summarized in Fig. 5.1 that influence network planning decisions of *service-drive networks*. Our model is able to account for all such relationships, including those that, as we will see in Sec. 5.5, are rather very weak.

System elements

Our system model includes four elements, represented by grey vertical blocks in Fig. 5.2: technologies, locations, user clusters and content types. The sets of existing technologies, locations, clusters and content types are, from the point of view of our system model, input data. In Sec. 5.3 and Sec. 5.4 we will discuss how and from which sources this information is gathered.

Technologies $t \in \mathcal{T}$ represent the available types of network infrastructure. LTE macro and micro base stations as well as WiFi and mmWave access points correspond to different technologies; furthermore, base stations using different frequencies or power levels also correspond to different technologies. As a general rule, if two infrastructure elements have different cost or coverage or performance, then in our model they correspond to different technologies. Some technologies require specific permissions or a license to operate (e.g. LTE base stations in licensed bands) and they are unlikely to be deployed by the OTTs. Therefore, we denote as $\mathcal{T}_{\text{OTT}} \subseteq \mathcal{T}$ the set of technologies available to the OTTs, while $\mathcal{T}_{\text{MNO}} \equiv \mathcal{T}$ corresponds to the technologies available to the MNOs.

Locations $l \in \mathcal{L}_t$ represent the positions in space at which infrastructure of technology $t \in \mathcal{T}$ may be located. As an example, each building within an urban area may correspond to a location. In the following, we will often refer to the combination of a location $l \in \mathcal{L}_t$ and a base station type $t \in \mathcal{T}$ as a *base station*.

User clusters $u \in \mathcal{U}$ represent groups of users that can be seen as co-located. Indeed, when performing network planning, we are not interested in the position or mobility of individual users, but rather in the *total* number of users in a constrained geographic area.

Finally, *Content types* $c \in \mathcal{C}$ (hereinafter “contents” for brevity) represent the types of content users are interested in accessing such as, for example, High-Definition (HD) movies or on-line gaming.

Parameters

Parameters are known quantities associated with one or more elements of our system model. They are represented by green horizontal blocks in Fig. 5.2. From the viewpoint of our model, they are input values; however, as denoted by the yellow clouds in the figure, they actually come from the sources listed in Sec. 5.3.

The first parameter is the *estimated capacity* $k(l, t, u)$ that a base station of technology $t \in \mathcal{T}$, built in location $l \in \mathcal{L}_t$, would be able to offer to users in cluster $u \in \mathcal{U}$ if it serves no other clusters and it is based on a simplified model. For example, $k(l, t, u) = f(t)B(t)\eta(l, t, u)$, where $f(t)$ is the fungibility of technology t as expressed in Sec. 5.3.2, $B(t)$ is the bandwidth available to technology t, and $\eta(l, t, u)$ is the spectral efficiency that base station of technology t in location l can deliver to user u. The spectral efficiency is estimated based on the distance between location and user cluster, the propagation model we assume, and the specific technology employed.

Furthermore, we need to know the *cost* $p(l, t)$ of building a base station of technology $t \in \mathcal{T}$ in location $l \in \mathcal{L}_t$. Costs ranges for different technologies can be extracted from the literature, as detailed in Sec. 5.3.

Last, we have the *demand* $\tau(c, u)$ requested by users in cluster $u \in \mathcal{U}$ for content type $c \in \mathcal{C}$.

Variables

Variables corresponds to the decisions MNO or OTTs can make. They are represented by blue horizontal boxes in Fig. 5.2.

The first task the MNO has face is to propose an SLA contract to the OTTs, where the OTTs has to pay for their contents to be given guaranteed bitrate. Such a fee may be a direct payment or some other form of revenue transfer from OTTs to the MNOs [43, 44]. In our model, we represent it through a per-megabit *fee* β charged to OTT to have its traffic served by the MNO's network.

The following decisions concern whether or not the OTTs *deploy* a base station of technology $t \in \mathcal{T}$ at location $l \in \mathcal{L}_t$ represented through a binary variable $y_{\text{OTT}}(l, t)$. Parallel decisions are how to serve the users, i.e., the fraction of the total time and frequency resources (in LTE terminology, physical resource blocks, PRBs) a base station of technology $t \in \mathcal{T}$, deployed at location $l \in \mathcal{L}_t$, uses to meet the demand of subscriber requesting content $c \in \mathcal{C}$ located in cluster $u \in \mathcal{U}$. This is expressed through a real-valued variable $x_{\text{OTT}}(c, l, t, u) \in [0, 1]$. Finally, the MNO has to deploy its own infrastructure and to decide how to serve the residual demand for all the contents. These decisions can be represented by $y_{\text{MNO}}(l, t)$ and $x_{\text{MNO}}(c, l, t, u)$ respectively. y_{MNO} and x_{MNO} have the same structure as y_{OTT} and x_{OTT} described earlier.

5.2 Solution concept

As mentioned earlier, our model accounts for the two main actors involved in deploying and managing service-driven networks, i.e., traditional MNOs and OTTs. Each actor is self-interested and ultimately aims at maximizing its own profit. In this section, we detail the joint decision process they take part in, and how individual decisions are made.

5.2.1 Decision process

In our model, both OTT service providers and the MNO seek to maximize their profit (or, equivalently, minimize their costs, as revenue obtained from end users is assumed constant). Both need to decide what infrastructure of their own to deploy, in which location, and of what type. OTT providers also need to decide how much to rely on the MNO to serve their content types, and MNO decides how much to charge OTTs to satisfy QoS requirements specified in the Service Level Agreement (SLA) for the OTTs' contents.

In the first stage, the MNO decides the fee, i.e., the per-megabit yearly price that OTTs have to pay if they want their contents to be delivered at a certain bitrate. Fees have an effect on the revenue the MNO collects. Intuitively, setting low fees indicates potentially low revenue for the MNO, which has to serve more traffic (hence update its network) for little additional revenue. On the other hand, setting very high fees represents a stronger incentive for OTTs to deploy their own infrastructure, and therefore also affect the MNO revenue.

In the second stage, OTTs have to plan their infrastructure. For each part of the topology, they can choose between having the MNO serve their demand therein – and pay the fee – or serving the demand themselves, deploying their own base stations – and paying the related cost.

In the third stage, MNOs have to make decisions regarding deployment in their network. They know they have to serve all the demand left unserved by the OTTs in order to honor their commitments, by deploying the necessary infrastructure while keeping their costs at a minimum.

It is worth stressing that, at every stage of the solution, we assume that the demand *will* be served, i.e., that the dimensioning problem is feasible. This assumption reflects the widespread belief that it will be possible for cellular networks to cope with the challenge posed by the increase in data demand, and it is our task to seek the best way to do so.

5.2.2 Individual steps

In the following, we detail how the MNO and the OTTs make their decisions in each step of the process described in Fig. 5.4, specifically, the problem they seek to optimize and the method they employ to do it. We start by presenting the deployment decisions for each OTT, then the subsequent deployment decisions for the MNO. We reserve special attention to the problem of setting the fee at the end of this section.

OTT - minimize the costs to serve the demand

We now focus on the deployment decision problem from the point of view of the OTTs. We assume that each OTT knows the characteristics of its own demand, expressed as $\tau(\widehat{c}, u)$, where $\widehat{c}$ is the content type that belong to the OTT considered. The OTT can choose between serving the demand directly, deploying its own infrastructure, or using the MNO's infrastructure and paying the fee β. We assume that the β fee at this stage is known. Each OTT wants to maximize its profit, which is in this case equivalent to minimizing the total cost, i.e.,

$$\min_{x_{\mathrm{OTT}}, y_{\mathrm{OTT}}} \left(\beta \sum_{u \in \mathcal{U}} \bar{\tau}(\widehat{c}, u) + \sum_{t \in \mathcal{T}_{\mathrm{OTT}}} \sum_{l \in \mathcal{L}_t} y_{\mathrm{OTT}}(l,t) p(l,t) \right) \tag{5.1}$$

$$\text{s.t.} \quad \sum_{t \in \mathcal{T}_{\mathrm{OTT}}} \sum_{l \in \mathcal{L}_t} x_{\mathrm{OTT}}(\widehat{c}, l, t, u) k(l,t,u) \leq \tau(\widehat{c}, u), \quad \forall u \in \mathcal{U} \tag{5.2}$$

$$\sum_{u \in \mathcal{U}} x_{\mathrm{OTT}}(\widehat{c}, l, t, u) \leq y_{\mathrm{OTT}}(l,t), \quad \forall t \in \mathcal{T}_{\mathrm{OTT}}, l \in \mathcal{L}_t. \tag{5.3}$$

The first term in Eq. (5.1) represents the fees paid to the MNO to serve the residual demand. The second term is the cost incurred by the OTT to deploy its own infrastructure. The OTT has no constraints to serve all its demand itself as indicated in Eq. (5.2): any residual demand $\bar{\tau}$ will be served by the MNO, as described earlier, in exchange for a fee. Moreover, constraint Eq. (5.2) forbids the OTT to serve more traffic demand then it has to, and it prevents the quantity $\bar{\tau}(\widehat{c}, u)$ from taking negative values. Eq. (5.3) ensures that we properly account for the maximum capacity of the base stations and that only active base stations, i.e., base stations for which the y_{OTT}-value is set to one, are used. If the y_{OTT}-value is zero, Eq. (5.3) indicates that the base station cannot serve any user at all. If it is one, it avoids that base stations are designated to serve more traffic than they can, i.e., more than their capacity. The OTT is concerned about minimizing its total cost expressed by Eq. (5.1); if building no base station at all serves such a purpose, there is nothing forcing OTTs to do otherwise.

Next, we have to set $\bar{\tau}$, that is the difference between the total demand and the demand that is served by the OTT's base stations:

$$\bar{\tau}(\widehat{c}, u) = \tau(\widehat{c}, u) - \sum_{t \in \mathcal{T}_{\mathrm{OTT}}} \sum_{l \in \mathcal{L}_t} x_{\mathrm{OTT}}(\widehat{c}, l, t, u) k(l, t, u). \tag{5.4}$$

It is worth to observe that $\bar{\tau}$ is a decision variable for the OTTs, and, after we obtain the $\bar{\tau}$ for all the content types $c \in \mathcal{C}$, it becomes an input parameter in the MNO deployment problem we examine in the next section.

MNO - minimize the deployment costs

At this stage, the MNO obtains the *residual* demand $\bar{\tau}(c, u)$ for each user cluster and content type, i.e., the demand that the OTTs decided to have served by the MNO, in exchange for the fee β. The MNO has to deploy infrastructure, i.e., setting y_{MNO}-values to one, in order to serve the residual demand. It seeks the deployment decisions that lead to demand satisfaction with minimum costs:

$$\min_{x_{\mathrm{MNO}}, y_{\mathrm{MNO}}} \sum_{t \in \mathcal{T}_{\mathrm{MNO}}} \sum_{l \in \mathcal{L}_t} y_{\mathrm{MNO}}(l, t) p(l, t). \tag{5.5}$$

$$\text{s.t.:} \quad \sum_{t \in \mathcal{T}_{\mathrm{MNO}}} \sum_{l \in \mathcal{L}_t} x_{\mathrm{MNO}}(c, l, t, u) k(l, t, u) \geq \bar{\tau}(c, u), \quad \forall c \in \mathcal{C}, u \in \mathcal{U} \tag{5.6}$$

$$\sum_{u \in \mathcal{U}} \sum_{c \in \mathcal{C}} x_{\mathrm{MNO}}(c, l, t, u) \leq y_{\mathrm{MNO}}(l, t), \quad \forall t \in \mathcal{T}_{\mathrm{MNO}}, l \in \mathcal{L}_t. \tag{5.7}$$

Eq. (5.5) has to satisfy the constraints on the traffic demand and on the capacity expressed by Eq. (5.6) and Eq. (5.7). The constraint expressed by Eq. (5.6) ensures that the MNO provides enough capacity for all user clusters $u \in \mathcal{U}$ and contents $c \in \mathcal{C}$ the MNO must serve, while the constraint in Eq. (5.7) ensures that we properly account for the maximum capacity of the base stations and that only active base stations are used.

We note that the fee β does not appear in the MNO deployment problem – as at this stage, the MNO has already established the fee, and has to serve all the traffic the OTT decides to delegate to it.

MNO - maximize the revenue by setting the fee β

Setting the fee is the most complex task. It is a decision taken by the MNO that depends on the response of the OTTs. The objective of the MNO is maximizing its own profit as shown in the following formula:

$$\max_{\beta}\left(\beta\sum_{u\in\mathcal{U}}\sum_{c\in\mathcal{C}}\bar{\tau}(c,u)-\sum_{t\in\mathcal{T}_{\mathrm{MNO}}}\sum_{l\in\mathcal{L}_t}y_{\mathrm{MNO}}(l,t)p(l,t)\right) \tag{5.8}$$

The objective function expressed in Eq. (5.8) is composed by two terms. The first on the left-hand side indicates the revenue that the MNO collects by serving the residual demand for each content type that belong to the OTTs. The second term on the right-hand side is the cost sustained by the MNO to deploy additional infrastructure to serve the residual demand. The residual demand appears twice in this optimization problem, explicitly in the objective function in the first term, and implicitly in the second term in the form of the constraint Eq. (5.6).

Unfortunately, the residual demand $\bar{\tau}(c,u)$ depends upon decisions taken by the OTTs that are affected by the fee β as we have seen previously. It is clear at this stage that the problem of optimizing the fees by the MNO is entwined with the problem faced by the OTTs to minimize their own costs and we cannot give a closed-form expression. To circumvent this issue, we depict a different strategy that we illustrate in Fig. 5.3.

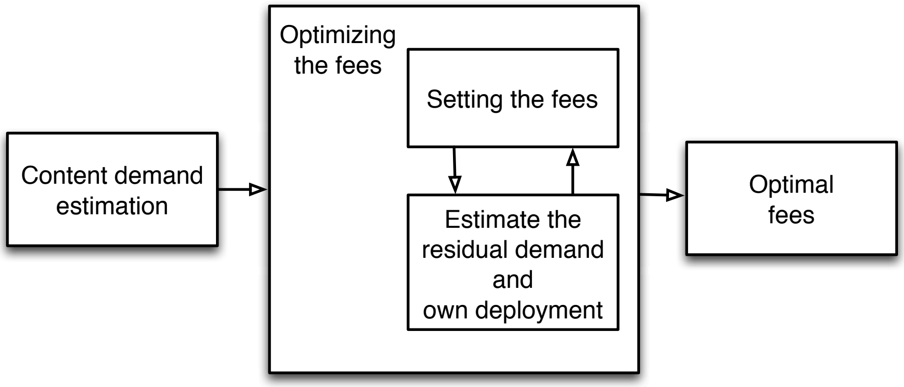

Content demand estimation
Optimizing the fees
Setting the fees
Estimate the residual demand and own deployment
Optimal fees

Figure 5.3: Strategy to set the optimal fees from the MNO perspective. The MNO has to *estimate* the action of each OTTs assuming that they are rational entities and that they will always take the possible best actions.

At first, the MNO observes the traffic demand and infers the characteristics of the demand. The MNO then tries to optimize the fees by evaluating its utility function expressed in Eq. (5.8) for several values of β and then selecting the best one. The MNO essentially has to solve a univariate discrete optimization problem that now involves an iterative process. The MNO *estimates* the residual demand left unserved by each OTTs and subsequently the new infrastructure necessary to serve the whole residual demand for each value of β he tries by sequentially solving the OTTs deployment and its own deployment. In order to correctly estimate the deployment, we assume that both the OTTs and the MNO seek always their optimal network configuration.

Each fee's configuration in fact yields to a different residual demand $\bar{\tau}$. However, the final network deployment corresponds to the one resulting from the β optimum from the perspective of the MNO.

5.2.3 Solution strategy

In the following, we discuss how each of the problems presented above can be solved.

The OTTs and MNO deployments, corresponding to the internal boxes in Fig. 5.4, have the same structure. They are linear problems with integer (binary) variables; they can be solved to optimality with branch-and-cut algorithms using commercial state-of-the-art solvers such as Gurobi or CPLEX, provided that the size of the problem is not too large.[2] If the size of the problem is too large, we would face the well-known scalability issues of problems belonging to the mixed integer linear programming (MILP) class. In this case, we would turn to heuristic solutions such as the ones described in [108], originally formulated for set-covering problems. Their greedy approach has been shown, through extensive studies carried out on the OR problem library [109], to consistently perform close to the optimum, with a ratio between the solution they find and the optimal one being typically around 1.2.

The problem solved by the MNO when setting the fee β is even more challenging, lacking a closed-form expression. However, in light of our new strategy described in Fig. 5.3, the new problem becomes a *univariate* problem in β, and therefore can be solved with root-finding algorithms such as the Brent method [110].

Root-finding algorithms explore several possible values of the decision variable, evaluate the value of the objective function, and use such information to select the next values to try. In our case, each iteration of the Brent method means solving the problems Eq. (5.1) and Eq. (5.5) in sequence; their outputs are used to compute the payoff Eq. (5.8) and hence to find the optimal value of β.

The scalability of the overall solution concept is ensured by the fact that the Brent method requires a limited number of iterations to converge, and each iteration takes a limited amount of time to solve subproblems Eq. (5.1) and Eq. (5.5).

Although they have been consistently observed to perform very well in practice, neither the Brent method nor the heuristic in [108] come with a formal, absolute optimality guarantee. This is indeed consistent with our goals: we seek to confirm and uncover correlations between the conditions under which service-driven networks will operate and the features they will exhibit; to this end,

[2] In our case we rely on the Gurobi solver [107].

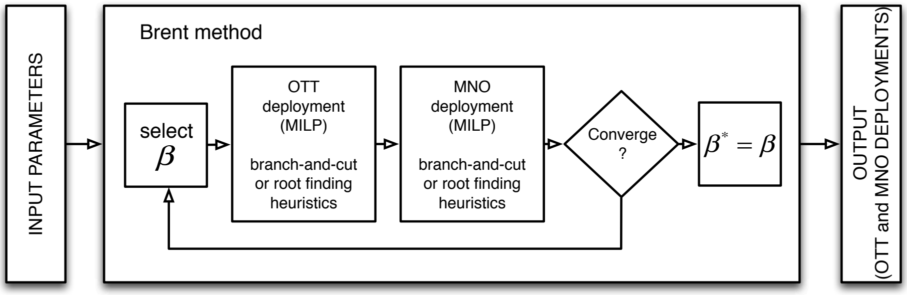

INPUT PARAMETERS
Brent method
select β
OTT deployment (MILP)
branch-and-cut or root finding heuristics
MNO deployment (MILP)
branch-and-cut or root finding heuristics
Converge ?
$\beta^* = \beta$
OUTPUT (OTT and MNO DEPLOYMENTS)

Figure 5.4: Solution strategy implementation overview.

near-optimal solutions are essentially as useful as optimal ones.

5.3 Factors shaping service-driven mobile networks

In this section, we describe the factors that will drive and shape service-driven mobile networks: the availability of new types of base stations; the fungibility of different portions of the spectrum; the regulator decisions constraining the deployment of network infrastructure; the demand it will need to serve.

For each of these elements, we explain how it is captured within the model described in Sec. 5.1. We then review the estimates for its value existing in the literature, and determine either a value or a range of values to use in our performance evaluation.

5.3.1 Base station technologies

Heterogeneity will be an important feature of service-driven mobile networks. Different types of base stations will coexist therein, including:

- LTE macro-base stations (macroBSs);
- LTE micro-base stations (microBSs);
- millimeter-wave base stations (mmWave);
- Wi-Fi access points;

Additional types of base stations can be added to $\mathcal{T}$ as sufficiently detailed information about them becomes available such as Google's Project Loon [111] LTE balloon-powered platforms operating in unlicensed bands used to provide LTE coverage to rural areas, or Facebook's Connectivity Project [112].

Infrastructure types that operate in licensed bands (hence with exclusive usage rights) can only be deployed by mobile networks operators (i.e., LTE macroBSs and LTE microBSs at 1.8GHz and

Table 5.1: Infrastructure types populating set $\mathcal{T}$.

Base station type	Frequency	Bandwidth	Max. spectral efficiency	Max. capacity	Tx power	Cost range	Approx. range
LTE macroBS	1.8 GHz (licensed)	20 MHz	4.4 bps/Hz [100]	88 Mbps	40 dBm	[10000,60000] €/year [59, 60]	Several hundreds of meters
LTE microBS-L	2.6 GHz (licensed)	20 MHz	4.4 bps/Hz [100]	88 Mbps	33 dBm	[2000,10000] €/year [59, 60]	Few hundreds of meters
LTE microBS-S	3.5 GHz (shared)	20 MHz	4.4 bps/Hz [100]	<88 Mbps (fungibility<1)	33 dBm	[2000,10000] €/year [59, 60]	Few hundreds of meters
mmWave [113, 114]	73 GHz (shared)	500 MHz	2.25 bps/Hz	<1125 Mbps (fungibility<1)	30 dBm	[1000,2000] €/year	Tens of meters
Wi-Fi	2.4,5 GHz (shared)	20 MHz	3.12 bps/Hz [115]	<62.4 Mbps (fungibility<1)	24 dBm	1000 €/year [59, 60]	Tens of meters

2.6GHz), while technologies that can operate in shared bands can be deployed by both MNO and OTT (i.e., mmWave, LTE microBSs at 3.5GHz, Wi-Fi) having different costs and fungibility (see Sec. 5.3.2).

Table 5.1 summarizes the types of infrastructure, i.e., the elements of set $\mathcal{T}$. For each of them, we indicate the frequency they operate at, their bandwidth and spectral efficiency, and the resulting maximum capacity – this is an upper bound on the values of parameter $k(l, t, u)$. Notice that microBSs with exclusive and opportunistic access are considered as two separate elements of $\mathcal{T}$.

Estimating the cost of a base station is a difficult exercise. We gathered the figures indicated in Table 5.1 preferring peer-reviewed publications where available, and falling back to other sources such as business/technical reports when needed. Despite extensive research, we were unable to find a single cost estimate for mmWave base stations, other than generic claims that they will be cheap. We conjecture that their cost will lie between the most expensive Wi-Fi access points and the cheapest microBSs.

Also notice that, for simplicity, we assume the price for each base station technology $t \in \mathcal{T}$ to be the same regardless of the location $l \in \mathcal{L}_t$ where they are built.

5.3.2 Spectrum fungibility

Fungibility as a concept has roots in the economic process of trade. The core of this concept is the ability to substitute one item with another, without altering any other aspect of a trade. As such, this concept implies two distinct roles, a buyer and a seller, with the former having the power to declare items fungible. In terms of spectrum, the buyer may be a network operator and the seller may be a regulatory agencies, offer usage of a particular set of frequencies. Two bands of spectrum may only be considered fungible if the operator is happy to obtain either one at a given time and price.

As more spectrum usage becomes more fluid, the concept of fungibility of spectrum becomes an increasingly important issue [38]. The increase in spectrum trading that results from the ability and motivation to rapidly change operating frequencies makes fungibility an important consideration for future radio systems. Network operators now have the ability to divert traffic between a number of different bands and technologies, buying this capacity on demand; fungibility provides an important tool in assessing the relative usefulness of these options based on the goals of the operator.

Different portions of spectrum are not fungible, for three orders of reasons:

- different frequencies are associated with different bitrate and coverage;
- different frequencies are typically used by different hardware, running different protocols;
- unlicensed frequency bands are more prone to congestion and interference than licensed ones.

In the context of wireless networking, *fungibility analysis* specifically refers to the last item: the purpose is to quantify the performance penalty incurred in by using unlicensed frequencies.

From the point of view of our model, all three aspects above are embedded in the capacity parameter $k(l, t, u)$, as discussed in Sec. 5.3.1. Fungibility is a coefficient by which we further multiply the capacity as listed in Table 5.1; its values are presented in Table 5.2.

Table 5.2: Fungibility coefficients.

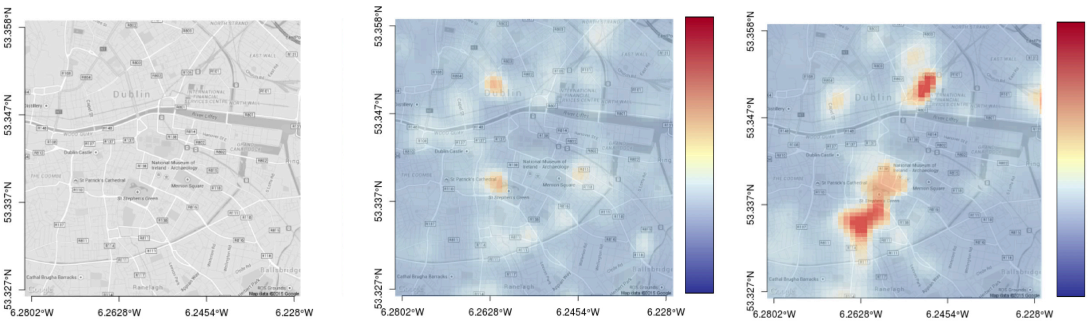
Technology	Fungibility range	Source
Small cells (WiFi and mmWave)	$0.6 - 0.9$	[116]
MicroBSs (in shared bands)	$0.5 - 0.75$	[117]

5.3.3 Regulatory decisions

Licensed spectrum is a scarce resource. Hence, who will be allowed to use it and how has a critical impact on network performance, and heavily depends on regulatory decisions. In the context of our work, we examine two scenarios concerning which spectrum, and under which conditions, OTTs have access to:

- LTE-standard complying base stations can only be deployed in licensed spectrum. OTTs cannot build microBSs in shared bands [Scenario A].
- LTE-standard complying base stations can be deployed in both licensed and shared spectrum. OTT and MNO can deploy microBSs, but in shared bands they experience a fungibility coefficient lower than one, as reported in Table 5.2 [Scenario B].

Selecting one or another of these scenarios changes the elements of set $\mathcal{T}_{\mathrm{OTT}}$, as well as the adjusted capacity k of base stations according their fungibility.



(a) (b) (c)

Figure 5.5: (a) reference topology; demand for content type $\widehat{c}$ for (b) low and (c) high complementarity.

5.3.4 Demand

The total demand and set $\mathcal{C}$ of content types depend on the reference scenario we consider, as discussed in Sec. 5.4 later. We seek to study the extent to which demand is *location-specific*, i.e., whether users requesting a particular content tend to be close to each other or not. We quantify this through the Hegyi index [118], used to express the strength of clustering between spatially distributed points (user clusters, in our case) associated to a continuous value (contents, in our case). The Hegyi index for user cluster u and content c is:

$$H(c,u)=\sum_{i=1}^{N}\frac{\tau(c,u)}{\tau(c,u)(1+||u+n_i(u)||)}, \tag{5.9}$$

where $||\cdot||$ denotes the Euclidean distance between two user clusters, $n_i(u)$ denotes the i-th closest cluster to u, and N is a parameter denoting the number of nearest user clusters we account for (in our case, $N=5$). The complementarity value associated with a specific content type $c'\in\mathcal{C}$ is defined as the average over all user clusters $u\in\mathcal{U}$ of $H(c',u)$:

$$H(c')=\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}H(c',u). \tag{5.10}$$

The complementarity value defined in Eq. (5.10) ultimately tells us how clustered demand for a certain provider types tends to be. Intuitively, we can expect that a more clustered demand is easier to serve through such targeted infrastructure as the one that can be deployed by OTTs. We correlate the complementarity of the demand with the traits of the resulting network in Sec. 5.5.

5.4 Reference scenario

In this section, we describe the reference scenario we employ for our simulations, i.e., how we populate the sets of user clusters $\mathcal{U}$ and (potential) base station locations $\mathcal{L}_t$ for each technology $t \in \mathcal{T}$, and how we set the demand $\tau(c, u)$ for each user cluster and content type.

User clusters and locations Our reference topology is the entire urban area of Dublin, Ireland, as indicated by the Central Statistics Office (CSO) Ireland in [99]; a section thereof is shown in Fig. 5.5(a). Using the same methodology explained in Sec. 3.2 we place a total of $2,210$ user clusters throughout this area, in such a way that each cluster represents a population of at most 300 people. It follows that more populated areas tend to have more user clusters; this enables us to study dense deployments in such areas while keeping the overall complexity low. We also assume that the LTE macroBSs deployment is already inplace to ensure full coverage and mobility and it is given by the data available on-line [67]. In fact it makes sense for the MNOs to consider in network planning expansion problems the infrastructure already deployed, especially the costly ones.

The possible locations in $\mathcal{L}_t$ for technology $t \in \mathcal{T}$ are uniformly placed on a regular grid with different inter-site distance depending on the coverage range. For example, inter-site distance for LTE microBS is 100 meters, while for WiFi and mmWaves is 50 meters.[3]

Service providers demand Setting the demand values $\tau(c, u)$ is a complex task, for which little information is available and some speculation is unavoidable. We proceed in three steps, as follows:

1. we set the total demand for each cluster u, i.e.,
   $\sum_{c \in \mathcal{C}} \tau(c, u)$;
2. we decide how this total demand is split between contents;
3. we adjust the resulting demand complementarity.

We accomplish the first step by leveraging a set of real-world call-detail record (CDR) information from an Irish mobile operator, referring to a period of two weeks in 2013. We augment that total demand according to the projections of the Cisco Virtual Network Index [94], and obtain an estimate for mobile data demand in the time of service-driven networks.

Breaking down such an aggregated demand into individual demand for each content is another complex problem. We turn to the measurement work [79], which identifies four *traffic patterns*, i.e., sets of content types, that users in each location were found to conform to. Each pattern includes

[3] Candidate locations are are commonly placed in regular grids in planning problems [58].

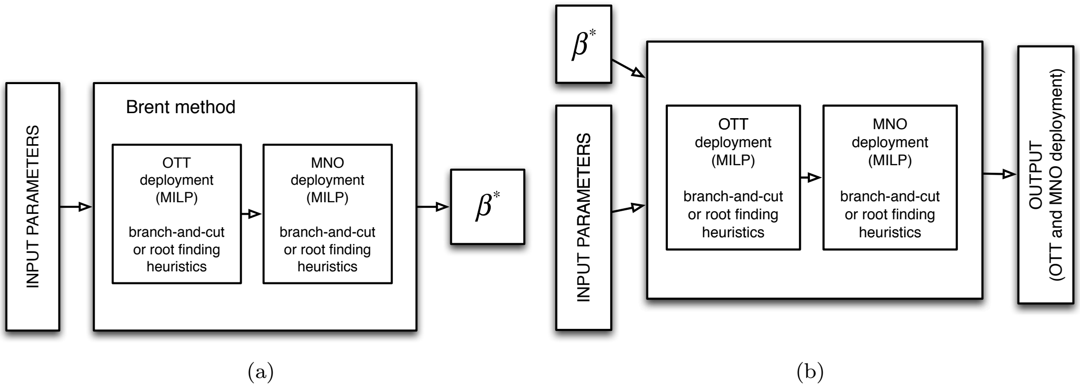

INPUT PARAMETERS
Brent method
OTT deployment (MILP) branch-and-cut or root finding heuristics
MNO deployment (MILP) branch-and-cut or root finding heuristics
β*
(a)
β*
INPUT PARAMETERS
OTT deployment (MILP) branch-and-cut or root finding heuristics
MNO deployment (MILP) branch-and-cut or root finding heuristics
OUTPUT (OTT and MNO deployment)
(b)

Figure 5.6: Analysis methodology. Once the process in Fig. 5.4 is completed, we analyse the impact of the input parameters by looking at the problem from two different angles. In (a), we check the influence of the input parameters on the price imposed by the MNO on the OTT ($\beta^\star$). In (b), we seek to find out the influence of the same input parameters and the $\beta^\star$ (now considered as an input parameter) on the output deployment.

several content types, and the same content type can appear in multiple traffic patterns. We assume then that a particular content type can be associated with an OTT provider.

We randomly associate one traffic profile to each user cluster, adjusting the average distance between two user clusters belonging to the same profile.[4] Intuitively, a small distance means that users wanting the same content types tend to be located close together, hence a higher complementarity as defined in Eq. (5.10).

Finally, we select the most popular of contents types as $\widehat{c} \in \mathcal{C}$, i.e., the demand volume that the OTT has to serve (either through the MNO's network or through its own). All other content types are assumed to belong to the MNO. Fig. 5.5(b) and Fig. 5.5(c) show the demand for $\widehat{c}$ in different parts of the topology when the complementarity index is at its lowest and highest value, respectively.

5.5 Results

In this section we conduct a sensitivity analysis on the relationships existing between the factors of interest that we believe will have an effect on the deployment decisions of an MNO and an OTT entering into a mutual service level agreement. We split the analysis in two parts according to Fig. 5.6. In the first part, we aim to understand the effects of the independent variables on the fee $\beta^\star$ the MNO charges the OTT to carry its traffic, i.e., the value of β that maximizes the quantity in Eq. (5.8) as depicted in Fig. 5.6(a). Once the MNO sets $\beta^\star$, $\beta^\star$ becomes an input parameter and,

[4] More details on the procedure can be found in Appendix B.

following our solution strategy, we finally assess which are the most important factors on service-driven deployment decisions for both the MNO and the OTT as described in Fig. 5.6(b). We justify the assumption that $\beta^\star$ is an input parameter because it is a decision that MNO should take a priori rather than a posteriori.

5.5.1 What determines $\beta^\star$?

The first part of the analysis, as indicated in Fig. 5.6(a), seeks to assess the key determinants on the variations of $\beta^\star$. We carry out a sensitivity analysis employing a multivariable[5] ordinary least-squared (OLS) regression model, where the independent variables are all the ones on the left hand side of Fig. 5.1 and the dependent variable is $\beta^\star$.

The results of the regression model are summarized in Table 5.3. We can immediately notice that among the input parameters chosen to analyse $\beta^\star$, only three are actually statistically significant at a p-value of 0.01. First, both the regulator and the cost of LTE microBSs have a negative impact on the fee charged by the MNO. The (binary) variable modelling the regulator allows the OTT access to deploy LTE microBSs in shared spectrum. Longer range infrastructures appear to be very appealing to OTTs since they allow the OTTs to serve lower-demand subscribers with less infrastructure.. As a consequence, the MNO tries to discourage the OTT from deploying many LTE microBS by setting a lower price on the OTT traffic it can serve. Second, the demand complementarity has a not-obvious positive impact on the $\beta^\star$. When the majority of the demand is clustered, it can be served with fewer base stations. As a result, the MNO's best action is to set a high fee so as to ensure some gain at least in those regions where OTT does not find it convenient to deploy any infrastructure. This dual effect is also captured by the histograms in Fig. 5.7. Forbidding OTTs from deploying microBSs, i.e., moving from scenario B to scenario A as defined in Sec. 5.3.3, has the general effect of increasing the fee – intuitively, reducing the freedom of action of OTTs increases the fee they are required to pay.

For completeness, we also report in Table 5.3 the coefficient of determination, R^2. It measures how well the model captures the variation of the dependent variable [120] and it is defined as:

$$R^2 = 1 - \frac{\sum_i (s_i - \bar{s})^2}{\sum_i (s_i - f_i)^2}, \tag{5.11}$$

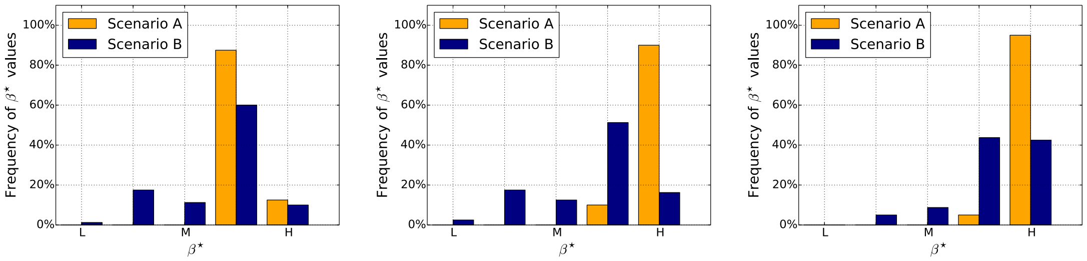
[5] There is often confusion in the regression analysis literature regarding the terminology used to describe different class of regression models. In this work we use the term *univariate* to refer to a model with a single dependent variable, *multivariate* to refer to a model with multiple dependent variables, and *multivariable* to refer to a model with multiple independent variables as explained in [119].

Table 5.3: Standardized regression analysis coefficients, dependent variable $\beta^\star$, $R^2 = 0.60$ (see Fig. 5.6(a)).

Independent variables	coef.*	p-val.**	std. error
Regulator	-0.55	✓	0.03
Demand complementarity	0.28	✓	0.02
Cost (mmwave)	-0.00	✗	0.09
Cost (micro LTE)	0.40	✓	0.04
Fungibility (small cells)	0.01	✗	0.10
Fungibility (micro LTE (S))	0.00	✗	0.12

* It represents the mean change in the dependent variable for 1 unit change in the corresponding independent variable while holding the other independent variables to their mean.

** ✓ indicates statistical significance (p-value$<$ 0.01) while ✗ indicates not statistically significant (p-value$>$ 0.01).



(a)

(b)

(c)

Figure 5.7: Medium infrastructure prices: distribution of the fee $\beta^\star$ for (a) low, (b) medium, and (c) high complementarity.

where s_i and $\bar{s}$ are the sample value and the sample mean respectively, and f_i is the modeled/fitted value.

5.5.2 What influences the deployment?

In the second part of the analysis, we focus on the network deployment. We identify, for each operator (i.e., MNO and OTT) and each technology (e.g. WiFi, mmWave, and LTE micro BSs) three parameters of interest: amount of infrastructure built, the capacity supplied and the traffic served by such infrastructure.

There are two main differences between the analysis in Sec. 5.5.1: first, the $\beta^\star$ identified in Sec. 5.5.1 takes now a pivotal role. Together with the original input parameters, we investigate the impact of $\beta^\star$ on the deployment, in particular the decision of the OTT of whether or not to deploy infrastructure to serve its own demand. To carry out this study we run multiple multivariable OLS regression in turn focusing on one outcome variable at a time. Other approaches could also be employed, such as multivariate analysis of variance (MANOVA) to explore the vector-space of the selected dependent variables. MANOVA is a powerful tool used extensively in regression analysis in social science: it is able to take into account multiple independent and multiple dependent variables within the same model, allowing a higher level of complexity. However, if the dependent variables are highly correlated, the chance that one of them becomes a near-linear combination of the other increases, leading the model to erroneous conclusions (this problem is known as *multicollinearity*).

In Table 5.4 we summarize the results obtained by checking one dependent variable at the time. Table 5.4 should be read by row, where each row reports the coefficients and p-value of a individual multivariable regression model on the corresponding network deployment output. Table 5.4 presents the list of network deployment quantities (i.e., number of base stations, capacity, traffic served, and level of clustering) grouped by operators and technology in the first column. Each of the remaining columns represents one independent variable (i.e., $\beta^\star$, regulatory decisions, demand characteristics, technologies cost, fungibility) and the final column report the R^2 value.

A few observations are noteworthy. First, only the cost for LTE microBSs is significant to all network planning actions. Second, a closer look at the upper part of Table 5.4 reveals that for the OTT, the $\beta^\star$ and the regulatory decisions are very influential parameters. They affect all the deployment decisions taken by the OTT, and, to a different extent, the MNO ones, i.e., defining the heterogeneity of the number of base stations and capacity deployed. The demand complementarity mainly impacts the planning decisions of the OTT, while the decisions taken by the MNO are mostly driven by the cost of infrastructure.

Infrastructure cost

As we have seen in Table 5.4, infrastructure cost has a significant influence on the deployment decisions taken by operators and service providers. In Fig. 5.8(a) we show the number of base stations of each type deployed by the OTT and the MNO. On the x-axis is the cost (relative to the WiFi) to deploy and maintain mmWave or an LTE microBS. OTTs are allowed to deploy LTE microBSs in the opportunistic access spectrum, i.e., we are in scenario B as described in Sec. 5.3.3.

Let us compare the group of bars on the left hand side of Fig. 5.8(a), indicating low costs for infrastructure with the group of bars in the middle and the right hand side of Fig. 5.8(a) indicating medium and high costs respectively: if infrastructure is sufficiently cheap, the best course of action for the MNO and OTT is to rely more on LTE microBSs, the ones that give the best compromise between coverage and capacity. As the infrastructure cost increases, both MNO and OTT rely more on small cells infrastructure. Fig. 5.8(b) shows how the capacity evolves according to changes in

Table 5.4: Standardized regression analysis coefficients, p-values and R^2 for the system in Fig. 5.6(b).

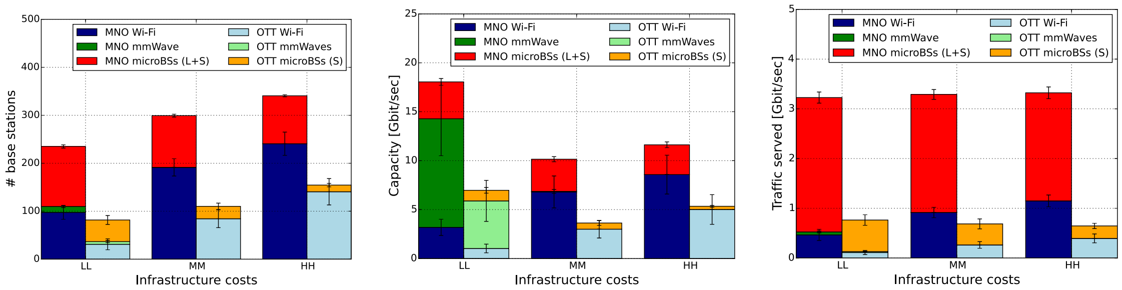
Dependent \ Independent			$\beta^\star$		**Regulator**		**Demand complementarity**		**Cost (mmWave)**		**Cost (micro LTE)**		**Fungibility (small cells)**		**Fungibility (micro LTE (S))**		$\mathbf{R^2}$
			coef.[*]	p-val.[**]	coef.	p-val.	coef.	p-val.	coef.	p-val.	coef.	p-val.	coef.	p-val.	coef.	p-val.	
OTT	WiFi	# bs	0.51	✓	−0.51	✓	−0.25	✓	0.09	✓	0.14	✓	−0.01	✗	−0.01	✗	0.90
		capacity	0.46	✓	−0.48	✓	−0.23	✓	0.13	✓	0.13	✓	0.31	✓	−0.01	✗	0.87
		traffic	0.48	✓	−0.48	✓	−0.33	✓	0.09	✓	0.18	✓	−26	✗	−53	✓	0.87
		clustering	0.54	✓	−0.38	✓	−0.15	✓	0.04	✓	0.25	✓	−0.00	✗	−0.02	✗	0.83
	mmWave	# bs	0.23	✓	−0.01	✗	−0.14	✓	−0.71	✓	0.06	✓	0.04	✓	−0.00	✗	0.75
		capacity	0.22	✓	−0.03	✗	−0.12	✓	−0.67	✓	0.05	✓	0.16	✓	−0.15	✓	0.71
		traffic	0.21	✓	−0.01	✗	−0.15	✓	−0.72	✓	0.07	✓	0.02	✗	−0.01	✗	0.76
		clustering	0.13	✓	−0.02	✗	0.04	✗	0.31	✓	0.07	✓	0.14	✗	−0.02	✗	0.20
	micro LTE	# bs	−0.23	✓	0.56	✓	0.11	✓	−0.01	✗	−0.34	✓	0.01	✗	−0.02	✗	0.69
		capacity	−0.23	✓	0.55	✓	0.11	✓	−0.02	✗	−0.33	✓	0.01	✗	0.11	✓	0.70
		traffic	−0.23	✓	0.62	✓	0.08	✓	−0.01	✗	−0.29	✓	0.01	✗	0.03	✗	0.81
		clustering	−0.25	✓	0.37	✓	0.01	✗	−0.01	✗	−0.32	✓	0.00	✗	−0.10	✓	0.51
MNO	WiFi	# bs	−0.05	✓	−0.78	✓	0.02	✓	0.10	✓	0.62	✓	0.00	✗	−0.05	✓	0.97
		capacity	−0.06	✓	−0.72	✓	0.02	✓	0.12	✓	0.57	✓	0.34	✓	−0.06	✓	0.94
		traffic	0.08	✓	−0.64	✓	0.05	✓	0.09	✓	0.73	✓	0.02	✓	−0.07	✓	0.96
		clustering	0.57	✓	−0.26	✓	−0.15	✓	0.05	✓	0.30	✓	−0.00	✗	−0.02	✗	0.78
	mmWave	# bs	0.01	✗	−0.21	✓	0.01	✗	−0.72	✓	0.20	✓	−0.00	✗	−0.01	✗	0.80
		capacity	0.02	✗	−0.20	✓	0.00	✗	−0.71	✓	0.19	✓	0.11	✓	−0.01	✗	0.78
		traffic	0.04	✗	−0.13	✓	−0.00	✗	−0.74	✓	0.20	✓	0.00	✗	−0.02	✗	0.82
		clustering	−0.14	✓	−0.38	✓	0.03	✗	−0.43	✓	0.29	✓	0.01	✗	0.00	✗	0.44
	micro LTE	# bs	−0.07	✓	0.87	✓	0.01	✓	−0.00	✗	−0.37	✓	−0.00	✗	−0.02	✓	0.98
		capacity	0.05	✓	0.86	✓	−0.00	✗	−0.00	✗	−0.47	✓	−0.06	✓	0.10	✓	0.96
		traffic	−0.03	✓	0.89	✓	0.01	✓	0.00	✗	−0.40	✓	−0.02	✓	0.04	✓	0.98
		clustering	0.13	✓	0.71	✓	−0.01	✗	0.01	✗	−0.61	✓	−0.02	✗	−0.01	✗	0.73

[*] It represents the mean change in the dependent variable for 1 unit of change in the corresponding independent variable while holding the others.

[**] ✓ indicates statistical significance (p-value<0.01) while ✗ indicates not statistically significant (p-value>0.01).

[***] We omit the to report the standard error and the 95% confidence interval in order to keep the table more readable.

[****] The colors are consistent with Fig. 5.8, Fig. 5.9, and Fig. 5.10 in order to help the reader to follow the discussion.



(a) (b) (c)

Figure 5.8: (a) number of base stations deployed, (b) capacity supplied, (c) traffic served for different combinations of prices. The error bars indicate the 95% confidence interval.

the costs. High-capacity/short-range technologies will only be successful if their cost is low enough; otherwise, due to their low coverage range, they are unlikely to be deployed. Looking at both Fig. 5.8(b) and Fig. 5.8(c) we discover an interesting effect: higher capacity does not necessarily translates into more traffic being served if the capacity is very localized, as it is the case with short-range infrastructures. In fact, as the costs for infrastructures increases, the OTT relies more on the MNO to serve its demand as it again can be observed in Fig. 5.8(c).

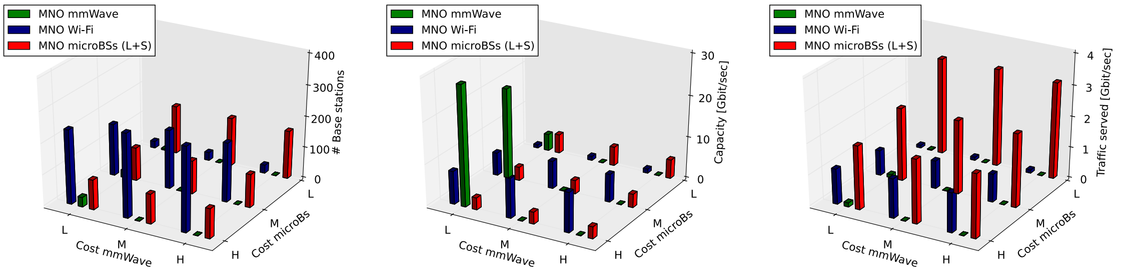
In Fig. 5.9 and Fig. 5.10 we give a closer look at how the combination of prices for microBSs and mmWave influence the decisions made by OTT and MNO. From Fig. 5.9(a) and Fig. 5.10(a) we can observe that OTTs essentially choose between microBSs and Wi-Fi access points, while they resort to mmWave base stations only when their cost low. Consistently with Fig. 5.8(b), Fig. 5.9(b) and Fig. 5.10(b) shows that mmWave base stations skew the network capacity, even when, as we can see in Fig. 5.8(c), Fig. 5.9(c) and Fig. 5.10(c), microBSs serve most of the traffic.

Service-driven networks will be different from current ones in that the network technology providing the most capacity may *not* be the one serving the most traffic. This leaves plenty of room for innovative applications such as proximity services and machine-to-machine systems – as long as they do not require ubiquitous, continuous coverage.

Demand complementarity

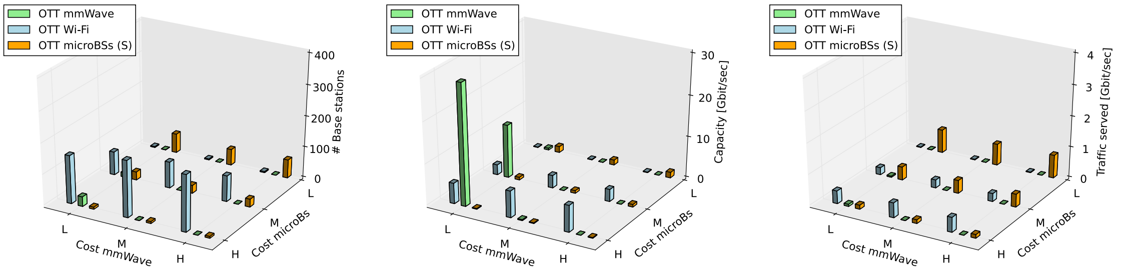
We now turn our attention to demand complementarity. As we have seen in Sec. 5.5.1 and in particular in Fig. 5.7, the demand complementarity has a not-negligible impact on the fee $\beta^\star$ that MNO charges the OTT to carry its traffic, i.e. the value of β that maximizes the quantity in Eq. (5.8).

Complementarity also affects the extent to which the network deployments of the MNO and, especially, of the OTT follow the demand. Fig. 5.11 refers to the scenario with medium infrastructure



(a) (b) (c)

Figure 5.9: MNO case. (a) number of base stations deployed, (b) capacity supplied, and (c) traffic served for different combinations of costs.



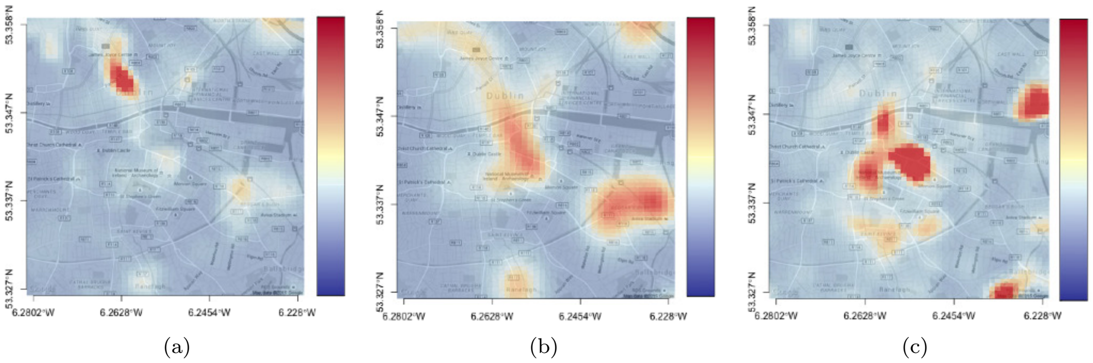
(a) (b) (c)

Figure 5.10: OTT case. (a) number of base stations deployed, (b) capacity supplied, and (c) traffic served for different combinations of costs.

prices and high complementarity. We can see from Fig. 5.11(a) that the demand for content $\widehat{c}$ tends to be very clustered. Fig. 5.11(b), showing the capacity deployed by the OTT, clearly follows the same pattern – the OTT deploys more capacity where it has more demand. The first thing we can observe by looking at Fig. 5.11(c), depicting the capacity deployed by the MNO, is that it does not clearly follow the demand of content $\widehat{c}$. Instead the capacity deployed in the reference scenario complements the one deployed by the MNO rather than overlapping with it.

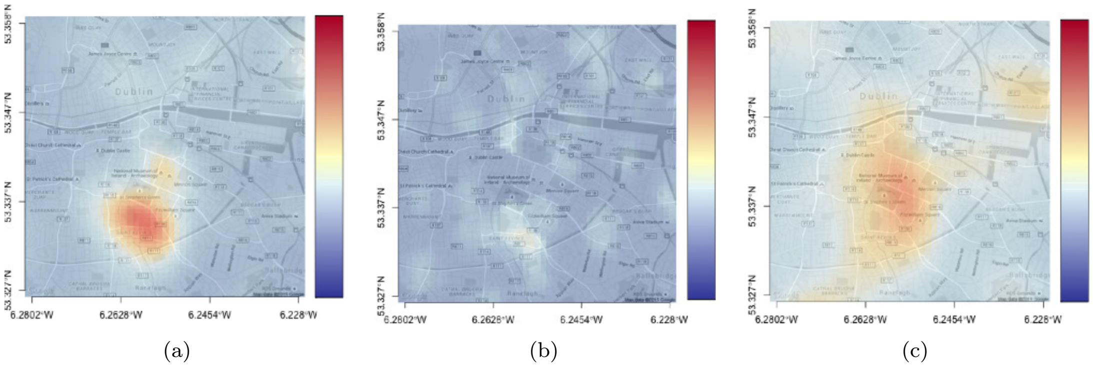
When we lower the complementarity to its minimum value, in Fig. 5.12, an entirely different picture emerges. The demand for content $\widehat{c}$ (Fig. 5.12(a)) is distributed over a wider area, mostly in the South. As we can see from Fig. 5.12(b), the OTT deploys a much lower number of base stations, in locations throughout the topology. Also notice from Fig. 5.12(c) how some of the demand for content $\widehat{c}$ is also served by the MNO.

In summary, Fig. 5.11 and Fig. 5.12, obtained for different values of complementarity, show us two altogether different networks. In Fig. 5.11, OTT and MNO complement each other in serving the traffic demand; in Fig. 5.12 the MNO builds a high-capacity network with vast coverage and the OTT confines itself to paying a (moderate) fee to use it.



(a) (b) (c)

Figure 5.11: High complementarity: (a) location of the demand for content type $\widehat{c}$; capacity deployed by (b) the OTT and (c) the MNO.



(a) (b) (c)

Figure 5.12: Low complementarity: (a) location of the demand for content type $\widehat{c}$; capacity deployed by (b) the OTT and (c) the MNO.

5.6 Conclusion

This chapter sought to study the mobile network expansion driven by the OTT's service needs and to assess, first, the factors that are likely to impact deployment decisions by OTTs and MNOs, and second, the characteristics of the resulting networks.

In Sec. 5.1, we introduced a model of service-driven networks accounting for their main features: heterogeneous infrastructure, working on both licensed and unlicensed frequencies; heterogeneous ownership, by both traditional mobile operators and OTT providers; a demand made up of location-specific, guaranteed-bitrate contents. In Sec. 5.2 we have discussed how MNOs and OTTs can efficiently make, self-interested, quasi-optimal decisions.

Sec. 5.3 contained the input information we use for our investigation. It is the result of an extensive search throughout scientific and economic literature and contains plausible ranges for the capacity, cost and coverage of the base stations that will be included within service-driven networks.

Using the data in Sec. 5.3 along the real-world scenario described in Sec. 3.2 we obtained the results presented in Sec. 5.5. We found that the factors with the deepest influence on the final network are not technical but rather economic (e.g., the cost of base stations) and regulatory (e.g., the type of spectrum and technologies OTTs are allowed to use). Different combinations of these factors yield radically different networks: in some cases, OTTs and MNOs deploy separate networks that complement each other; in others, MNOs find it convenient to offer plentiful, cheap capacity to OTTs, which have no incentive to deploy their own networks. Furthermore, the technologies providing most of the network capacity (most notably, mmWave) will not serve most of the traffic, leaving plenty of opportunities for innovative applications such as proximity services and M2M.

6 Spectrum Aggregation

A parallel and equally important trend to radio access infrastructure sharing in mobile networks is spectrum sharing. Conventionally, in mobile networks, spectrum is divided into frequency bands, with each band being exclusively reserved for use by a particular service, for example 3G or LTE. Each operator obtains frequency bands through a spectrum auction regulated internationally by International Telecommunication Union (ITU) and locally by national and regional agencies. From the perspective of an operator, exclusive access to spectrum allows the operator to have sole control over the offered quality of service, and it secures a return on large-scale infrastructure investments [8]. However, from a broader perspective, exclusively licensed spectrum remains under-utilized [80]. Given the low spatial and temporal correlation in mobile demand across mobile operators we discussed in Sec. 3.1, there are good reasons to seek more elastic models of spectrum usage, which is the central premise of much of the efforts put into dynamic spectrum access (DSA) research.

In this chapter we analyze the spectrum sharing problem from the perspective of competing mobile network operators. In particular, our approach relies on a key technique introduced in LTE-Advanced [37], Release 10, Carrier Aggregation (CA). CA provides the ability to aggregate contiguous and non-contiguous portions of the spectrum in order to sustain higher datarates. Arguably the most interesting feature of CA is that the component carrier to be aggregated may belong to non-contiguous bands, making it feasible for operators with licensed access to their spectrum, to share a portion of their exclusive bandwidths on an as-needed basis. In this particular setting, CA can be seen as a specific instance of spectrum sharing, which refers to the ability to re-use spectrum between operators whenever and wherever capacity expansion is needed. We extend the concept of CA by investigating the ability of independent mobile network operators (MNOs) to dynamically schedule access to portions of each others' spectrum.[1]

In our model, each MNO allows a portion of its exclusively licensed frequencies to be aggregated

[1] The model and results of this chapter have been presented at the IEEE International Conference on Communication (ICC), 2013. This work was done in collaboration with, and led by, Dr. Yong Xiao.

by other MNOs for a limited amount of time. We refer to this type of dynamic access as Dynamic Inter-network Carrier Aggregation (DI-CA) [121]. Essentially, DI-CA involves one MNO temporarily releasing some of its exclusive spectrum to another MNO, which can then be aggregated it with its own spectrum.

We propose a generalized model for the DI-CA-based system where an MNO can decide the amount of time to grant access to its licensed spectrum to other operators to maximize its payoff. In our generalized model, the payoff of each MNO can be any performance measure; however, we concentrate on the downlink channel capacity and we allow operators to access each other's spectrum for a certain amount of time, indicated as scheduling pair. We employ a methodology based on game theory different from the network optimization strategy adopted for Chapter 4 and Chapter 5, to address the following questions:

1. Under which conditions can DI-CA improve performance for all the MNOs?
2. How to achieve distributed scheduling when each MNO does not have any instantaneous information (e.g., payoffs, aggregated bandwidth) about others?

Answering the first question is important since it sets the conditions under which dynamic inter-spectrum aggregation takes place. To this end we assume that global information is available. After we derive the conditions under which dynamic inter-network carrier aggregation can improve the performance for all MNOs, we introduce a Nash Bargaining Solution-based fair spectrum scheduling pair scheme. Then, we devise a Bayesian Game where independent operators can decide whether or not they can increase their performance by sharing dynamically the spectrum without information on the other MNO. Finally, we propose a distributed algorithm that converges to a neighbourhood of the Bayesian Nash Equilibium (BNE).

6.1 Network Model

Consider a mobile cellular environment that consists of two MNOs that control the dynamic aggregation of their component carriers through two entities that we refer to as Mobile Network Aggregators (MNAs), M_1 and M_2. Each MNO operates base stations that we assume are co-located. Let the bandwidth under the control of the MNAs M_1 and M_2 be B_1 and B_2, respectively.

Let us now assume that the revenue each MNA obtains from aggregation is proportional to its transmit bandwidth. We can define the payoff/utility of M_i without DI-CA by $\varpi_i^{noCA} = B_i R_{ii}$ where R_{ii} is the revenue/benefit per Hertz obtained by M_i without DI-CA. In this chapter, we define the payoff of each MNA M_i as the downlink capacity, expressed as

$$R_{ii} = \mathbb{E}_{n \in \mathcal{K}_i}[log(1 + h_{ii}(n)w_i)] \tag{6.1}$$

where w_i is the transmit power of M_i, $\mathcal{K}_i$ is the set of subscribers served by M_i, and $h_{ii}(n)$ is the channel gain between MNA M_i and each mobile subscriber $n \in \mathcal{K}_i$.

Each MNA decides the assignment of resource blocks, i.e., the portion of the transmission time and bandwidth that is allowed to be aggregated by others. To ensure a minimum quality of service (QoS) for its own subscribers, each MNA can reserve a certain portion of the total bandwidth exclusively for its own transmission and only allows a portion of its frequency band to be aggregated by others using DI-CA. Let the reserved bandwidth and the dynamically aggregatable bandwidth of the MNA M_i be B_{ii} and B_{i-i}, respectively, where $-i$ denotes the MNA other than M_i, and

$$B_i = B_{ii} + B_{i-i}, \quad \forall i \in \{1, 2\}. \tag{6.2}$$

Note that the value of B_{ii} is used to maintain the basic performance for M_i and hence should be decided by the specific system requirements, for example, the long term average traffic, the minimum QoS, the expected number of subscribers. In addition, we assume B_{ii} and B_{i-i} to be fixed during the entire transmission process. We assume the revenue per Hertz of each MNO in different spectrum portions within its licensed bandwidth to be based on the same system requirements, and hence can write the revenue of M_i in the reserved bandwidth B_{ii} as

$$\pi_i^{CA1} = B_{ii} R_{ii}. \tag{6.3}$$

To start aggregating the bandwidth of others, the two MNAs will first exchange information regarding the time scheduling and aggregatable bandwidth of MNAs [122]. Assume each MNA M_i only allows the other MNA to schedule a proportion p_i of its transmission time dynamically aggregating its bandwidth B_{i-i}. Hence, the revenue that M_i obtains from its own subscribers in non-reserved frequency band B_{i-i} is given by

$$\pi_i^{CA2} = (1 - p_i) B_{i-i} R_{ii}. \tag{6.4}$$

Similarly, we can write the revenue that M_i obtains by dynamically aggregating the frequency band of M_{-i} for p_{-i} portion of time as

$$\pi_i^{CA3} = p_{-i} B_{-ii} R_{i-i} \tag{6.5}$$

where R_{i-i} is the revenue per Hertz when M_i aggregates B_{-ii} portion of the bandwidth of M_{-i}. If B_{-ii} and B_i are contiguous to each other, the revenues per Hertz R_{ii} and R_{i-i} may be similar to each other. However, if the frequency bands of M_i and M_{-i} are non-contiguous, R_{ii} and R_{i-i} are likely to be significantly different because of frequency selective fading and Doppler shifts in different frequency bands [123]. It is observed that CA introduces more complexity and cost in the system implementation even of the static kind that does not involve accessing CCs from other networks, i.e. CA requires more complex signalling and scheduling, and for non-contiguous CA, extra antenna/RF chains should be employed to allow each MNO to access the spectrum of others. In this model, we assume the extra cost brought by DI-CA for each MNO M_i is a constant denoted by ζ_i. We hence can write the payoff/utility of M_i when using DI-CA to be

$$\varpi_i^{CA} = \pi^{CA1} + \pi_i^{CA2} + \pi_i^{CA3} - \zeta_i. \tag{6.6}$$

6.2 Feasible Condition and Fair Scheduling

Both MNAs will independently decide whether or not to use DI-CA. This situation can be modeled using cooperative game theory, where autonomous decision makers (here, the MNAs) decide to cooperate if and only if cooperation brings mutual benefits [124] and the resource allocation is fair. In this section, we first derive the feasible condition for which DI-CA could provide payoff improvement for both MNAs, and then we briefly discuss the fairness criteria for the scheduling between MNAs. Let us first define the feasible DI-CA scheduling pairs for MNAs as follows.

Definition 2. A DI-CA scheduling pair (p_1, p_2) for MNAs is *feasible* if $0 \leq (p_1, p_2) \leq 1$, $\varpi_1^{CA}(p_1, p_2) \geq \varpi_1^{noCA}(p_1, p_2)$ and $\varpi_2^{CA}(p_1, p_2) \geq \varpi_2^{noCA}(p_1, p_2)$.

Following Definition 2, we have the following results about the feasible scheduling pairs for DI-CA.

Proposition 1. *There exists at least one feasible DI-CA scheduling pair* (p_1, p_2) *if*

$$\zeta_1 R_{22} + \zeta_2 R_{12} \leq B_{12}(R_{21} R_{12} - R_{11} R_{22}) \tag{6.7}$$

$$\zeta_1 R_{21} + \zeta_2 R_{11} \leq B_{21}(R_{21} R_{12} - R_{11} R_{22}) \tag{6.8}$$

are satisfied.

Proof:

Let us assume the condition in Eq. (6.7) is satisfied. In this case, we can claim that there exsists at least one pair of (p_1, p_2) for $0 < p_1 \leq 1$ and $0 < p_2 \leq 1$ which satisfies:

$$p_1 B_{12}(R_{12}R_{21} - R_{11}R_{22}) = \zeta_1 R_{22} + \zeta_2 R_{12}, \tag{6.9}$$

$$p_2 B_{21}(R_{12}R_{21} - R_{11}R_{22}) = \zeta_1 R_{11} + \zeta_2 R_{21}. \tag{6.10}$$

Let us rewrite Eq. (6.9) as follows,

$$\begin{aligned} p_1 B_{12}(R_{11}R_{22} - R_{21}R_{21}) + p_2 B_{21} R_{12} R_{22} - p_2 B_{21} R_{12} R_{22} = \zeta_1 R_{22} + \zeta_2 R_{12} \\ \Rightarrow R_{12}[p_1 B_{12} R_{21} - p_2 B_{21} R_{22} - \zeta_2] + R_{22}[p_2 B_{21} R_{12} - p_1 B_{12} R_{11} - \zeta_1] = 0 \\ \Rightarrow R_{12}\Delta\varpi_2 + R_{22}\Delta\varpi_1 = 0. \end{aligned} \tag{6.11}$$

Similarly, we can rewrite (6.10) as follows:

$$R_{11}\Delta\varpi_2 + R_{21}\Delta\varpi_1 = 0. \tag{6.12}$$

Combining Eq. (6.11) and Eq. (6.12) and using the fact that R_{11} (or equivalently R_{22}) and R_{12} (or equivalently R_{21}) are always positive and different from each other, we can claim that $\Delta\varpi_1 = 0$ and $\Delta\varpi_2 = 0$ must be simultaneously satisfied. In other words, (p_1, p_2) satisfies the condition of feasibility in Proposition 1.

■

Note that if the cost of using DI-CA is negligible, i.e., $\zeta_1 = \zeta_2 = 0$, then Eq. (6.7) and Eq. (6.8) turn into $R_{21}R_{12} \geq R_{11}R_{22}$. Proposition 1 provides a basic condition for which DI-CA can provide mutual benefits for both MNAs. However, it does not guarantee the spectrum sharing among MNAs to be fair. In fact, in cooperative Game Theory, several fairness criteria have been introduced to solve this problem, such as the Shapely value fairness [125] and the nucleolus fairness [126] which, however, are calculated in case the coalition is established. Another important problem for DI-CA is that there exists a fundamental tradeoff for each MNA M_i to choose p_i. More specifically, if one MNA M_i wants to reserve a large transmission time for its own transmission, i.e., choose a small

p_i, it will also decrease the chances of attracting other MNAs to share their spectrum. On the other hand, if p_i is large, it will decrease the capacity reserved for the MNA's own use. We then introduce the Nash bargaining solution (NBS) to address the fairness problem where the players seek to maximize the product of the following function:

$$\max_{p_1,p_2} \quad (\varpi_1^{CA} - \varpi_1^{noCA})(\varpi_2^{CA} - \varpi_2^{noCA}) \tag{6.13}$$

Intuitively, the solution of Eq. (6.13) (also referred as the Nash product [127]), denoted as (p_1^{NBS}, p_2^{NBS}), consists of each player calculating its non-cooperative payoff in addition to an equal share of the benefits occurring from cooperation. We will present the numerical results for NBS in Section 6.4. However, the NBS (as the Shapely value and the nucleolus fairness) requires each MNA to have global information, i.e., M_i needs to know R_{-ii} and $R{-}i{-}i$, to calculate (p_1^{NBS}, p_2^{NBS}). First of all, it may result in high communication overhead between the MNAs. The pair (p_1^{NBS}, p_2^{NBS}) also changes with the channel gains. Hence, whenever the channel condition changes, both MNAs will have to calculate these scheduling pairs increasing the computational complexity of DI-CA. Second, an MNA may not be willing to share all their private information with a competitor. In the next section, we propose a model-free Bayesian coalition formation framework which allows the MNAs to avoid exchanging this information when it is clear that DI-CA cannot improve their payoffs. This approach can solve the signalling, communications overhead and increased complexity issues introduced by DI-CA.

6.3 DI-CA with Incomplete Information

We assume the revenues of both MNAs change with time and the transmission process can be divided into N time slots, for each of which the revenues can be regarded as constants. We use $[m]$ to denote the transmission in the m^{th} time slot.

Definition 2: Let us define a Bayesian CA scheduling game as $G = \langle \mathcal{C}, a, \mathcal{T}, \pi, I \rangle$ where

- $\mathcal{C} = \{M_1, M_2\}$ is the set of players (MNAs),
- $\mathbf{a} = \{\mathbf{a}_1, \mathbf{a}_2\}$ where $\mathbf{a}_i = [a_i[1], a_i[2], \ldots, a_i[N]]$ and $a_i[m] \in \{0, 1\}$ is the action of player i; $a_i[m] = 0$ (or $a_i[m] = 1$) means that M_i does not (or does) use DI-CA,
- $\mathcal{T}$ is the set of types, which includes the instantaneous payoffs a player can achieve through its licensed and aggregated spectrum. Each MNA does not know the type of others,
- π_i is the instantaneous payoff of a player i.

There are generally two forms of uncertainty in a Bayesian coalition formation game: 1) type uncertainty: each player does not know the type of the other players, 2) sharing uncertainty: each player does not know how the resource will be shared among all the members after a coalition has been formed. In this work, we ignore the sharing uncertainty and assume that when a coalition has formed, players will exchange enough information and use NBS as the fairness criterion to share their own resources.

We focus on an infinite horizon, i.e. $N \to \infty$. Each MNA tries to maximise its average payoff,

$$\bar{\varpi}(\mathbf{a}_1, \mathbf{a}_2) = \mathbb{E}_m \varpi_i(R_{ii}[m]|p_i[m], p_{-i}[m], R_{i-i}[m], R_{-ii}[m], R_{-i-i}[m]). \tag{6.14}$$

We seek an equilibrium point of the system called the Bayesian Nash equilibrium (BNE). Note that, in our model, each player i cannot know the instantaneous payoff of the other players for the current time slot. Hence, at the beginning of a time slot m, player i will decide whether or not to form a coalition with others by using its private information and observation history. If the coalition formation request of player i has been rejected by others, $a_i[m] = 0$, player i will not exchange any information with others. However, if a coalition containing player i has been formed, i.e., $a_i[m] = 1$, all the coalition members can know the instantaneous information of each other and use NBS to share their resources.

We can rewrite the payoff of M_i in each time slot m as follows,

$$\varpi_i[m] = B_i R_{ii}[m] + a_i[m]\Delta\varpi_i[m] \tag{6.15}$$

where $\Delta\varpi[m] = p_{-i}^{NBS}[m]B_{-ii}R_{i-i}[m] - p_i^{NBS}[n]B_{i-i}R_{ii}[m] - \zeta_i$. It can be easily shown that the *best response* of M_i in each time slot m is given by

$$a_i[m] \begin{cases} 1, & \text{If} \quad \Delta\varpi_1[m] + \Delta\varpi_2[m] > 0 \\ 0, & \text{Otherwise.} \end{cases} \tag{6.16}$$

We assume that, at the beginning of each time slot m, each MNA i can only know $R_{-ii}[m]$. Because each MNA M_i cannot know $R_{-i-i}[m]$ and $R_{-ii}[m]$, let us define $\eta_i[m] = B_{i-i}R_{-ii}[m] + \frac{p_{-i}^{NBS}[m]B_{-ii}R_{i-i}[m] - p_{-i}^{NBS}[m]B_{-ii}R_{-ii}[m] - \zeta_i - \zeta_{-i}}{p_i^{NBS}[m]}$. We assume $\eta_i[m]$ is bounded and the condition for using DI-CA (i.e., $a_i[m] = 1$) in (Eq. (6.16)) can be rewritten as

$$\eta_i[m] > B_{i-i}R_{ii}[m] \tag{6.17}$$

The main focus here is to let each MNA M_i estimate whether or not (Eq. (6.17)) is satisfied by using the known $B_{i-i}R_{ii}[m]$ and an estimated version of $\eta_i[m]$, denoted by $\tilde{\eta}_i[m]$. In this work we use the average value as the estimate, i.e., $\tilde{\eta}_i = \mathbb{E}_{l=\{1,2,\ldots,m\}}\eta_i[l]$. More specifically, we assume each MNA can use the stochastic approximation method [128] to estimate the average value from its samples in an online fashion. Then, it is observed from Chebyshev's inequality, if the statistics of all the channels' gain are unchanged, nearly all the sample values are close to a neighbourhood of the mean.

Let us present the detailed algorithm below.

Algorithm 7 Distributed Scheduling Algorithm

```
1) Initialization: $m = 0$
    - $a_1[0] = a_2[0] = 1$,
    - Define a set of the tile slots $\mathcal{L} \subseteq \{1, 2, \ldots, N\}$ when both MNAs to use DI-CA, i.e., initially
      $\mathcal{L} = \emptyset$,
    - $1 \leq t \ll N$ is a pre-defined integer denoted as the training time duration.
2) Training: For $m = 1 : t$,
    a) $a_1[m] = a_2[m] = 1$,
    b) At the end of time slot $m$, each MNA knows $\pi^{CA}[m]$ and $\pi^{noCA}[m]$ and hence can calculate
       an estimate of $\eta_i[m]$, i.e., $\tilde{\eta}_i[m] = \frac{1}{m}\sum_{k=1}^{m}\eta_i[k]$. Both MNAs update $\mathcal{L} = \mathcal{L} \cup \{m\}$.
3) Decision Making: For $m = l + 1 : N$, at the beginning of every time slot $m$, $M_i$ observes
   $B_{i-i}R_{ii}[m]$ and knows $\tilde{\eta}_i[m]$,
    a) If $B_{i-i}R_{ii}[m] < \tilde{\eta}_i[m]$, $M_i$ sends a DI-CA request to $M_{-i}$. If $M_{-i}$ starts to communicate
       with $M_i$ to start using DI-CA during time slot $m$ and then goes to Step 4). Otherwise, $M_{-i}$
       sends a rejection message to $M_i$ and neither MNA can use DI-CA in time slot $m$. Repeat
       Step 3).
    b) If $B_{i-i}R_{ii}[m] \geq \tilde{\eta}_i[m]$, $M_i$ does not use DI-CA. Repeat Step 3).
4) Belief Update: If both MNAs use DI-CA in time slot $m$, each MNA $M_i$ updates its belief about
   $\eta_i[m]$ by
       $$\tilde{\eta}_i[m] = \tilde{\eta}_i[m-1] + |\mathcal{L}|\eta_i[m]/(|\mathcal{L}| + 1) \qquad (6.18)$$
   and both MNAs update $\mathcal{L} = \mathcal{L} \cup \{m\}$.
```

The main idea of the above algorithm is to let each MNA first estimate an initial value of $\tilde{\eta}_i$ using the training step, and then use the estimated $\tilde{\eta}_i$ to decide whether or not the condition in Proposition 1 is satisfied. Each MNA only sends the DI-CA request to the other when it decides that DI-CA can improve the performance of both MNAs. After each iteration, both MNAs update their beliefs about $\tilde{\eta}_i$.

We have the following result about the above algorithm.

Theorem 1. *If the expectation and variance of $\tilde{\eta}_i[m]$ are fixed during the transmission process and $\Delta\varpi_i[m]$ and $R_{ii}[m]$ are bounded, i.e., $0 \leq \Delta\varpi_i[m] \leq \Delta\varpi^+$ and $R_{ii}^- \leq R_{ii}[m] \leq R_{ii}^+$ $\forall m = 1, 2, \ldots, N$, then Alg. 7 converges to a BNE within a distance of $\epsilon(\epsilon-BNE)$ where $\epsilon = \frac{var(\eta_i)\Delta\varpi_i^+}{(B_{i-i}R_{ii}^- - \mathbb{E}\eta_i)^2}$.*

Proof:

As mentioned in Sec. 6.2, if both MNAs can have perfect information, i.e. each MNA knows $\eta_i[m+1]$ at the beginning of time slot $m+1$, it will use (or not use) CA if Eq. (6.16) is satisfied (or not satisfied). In other words, the optimal average payoff of M_i during the first $m+1$ time slots of transmission is achieved by allowing both MNAs to use CA when $\Delta\varpi_i[n] > 0$ and to stop using CA when $\Delta\varpi_i[n] \leq 0$ for all $n \in \{1, 2, \dots, m+1\}$.

Let us now consider the payoff achieved by Alg. 7. The main idea of Alg. 7 is to use the average value as the estimated version of the unknown $\eta_i[m]$. It has already been proved in [128] that the belief update in Step 4) of Alg. 7 always converges to the average value of η. Let us assume $\tilde{\eta}_i \approx \mathbb{E}_m(\eta_i[m])$. We then can focus on the convergence performance of the decision making process in Step 3). Let us consider the following cases in time slot m of Alg. 7,

$$\begin{aligned}
\bar{\varpi}_i[m+1] &= \frac{m\bar{\varpi}_i[m] + \varpi_i[m+1]}{m+1} \\
&\approx \bar{\varpi}_i + \frac{1}{m+1}[B_i R_{ii}[m+1] \\
+\mathbb{I}_{a_i[m+1]=1}&\mathbb{J}_{\eta_i[m+1]>B_{i-i}R_{ii}[m+1]}\Delta\varpi_i[m+1] \\
+\mathbb{I}_{a_i[m+1]=1}(1-&\mathbb{J}_{\eta_i[m+1]>B_{i-i}R_{ii}[m+1]})\Delta\varpi_i[m+1] \\
+(1-\mathbb{I}_{a_i[m+1]=1})&\mathbb{J}_{\eta_i[m+1]>B_{i-i}R_{ii}[m+1]}0 \\
+(1-\mathbb{I}_{a_i[m+1]=1})(1-&\mathbb{J}_{\eta_i[m+1]>B_{i-i}R_{ii}[m+1]})0],
\end{aligned} \tag{6.19}$$

where $\mathbb{I}$ and $\mathbb{J}$ are indicator functions. Note that $\Delta\varpi_i[m+1] < 0$ when $a_i[m+1] = 1$ and $\eta_i[m+1] < B_{i-i}R_{ii}[m+1]$.

Let us now prove that $\lim_{m\to\infty} \bar{\varpi}_i[m+1] \to \bar{\varpi}_i^\star \pm \epsilon$. Assuming the probability distribution function of η_i is fixed and using the fact that, in Alg. 7, $a_i[m+1] = 1$ when $\eta_i[m+1] > B_{i-i}R_{ii}[m+1]$, we can rewrite the above inequality as follows,

$$\begin{aligned}
&\lim_{m\to\infty} \frac{1}{m}\sum_{l=1}^{m} \varpi_i[l] - \varpi_i^\star \\
=& \lim_{m\to\infty} \frac{1}{m}\sum_{l=1}^{m}(\mathbb{I}_{a_i[l]=1}(1-\mathbb{J}_{\eta_i[l]>B_{i-i}R_{ii}[l]}) \\
&\Delta\varpi_i[l] + (1-\mathbb{I}_{a_i[l]=1})\mathbb{J}_{\eta_i[l]>B_{i-i}R_{ii}[l]}\Delta\varpi_i[l]) \\
\leq& \Pr(\eta_i > B_{i-i}R_{ii} \geq \mathbb{E}(\eta_i))\Delta\varpi_i^+ \\
&+\Pr(\eta_i \leq B_{i-i}R_{ii} < \mathbb{E}(\eta_i))\Delta\varpi_i^+ \\
=& \Pr(|\mathbb{E}(\eta_i) - B_{i-i}R_{ii}| < |\eta_i - \mathbb{E}(\eta_i)|)\Delta\varpi_i^+.
\end{aligned} \tag{6.20}$$

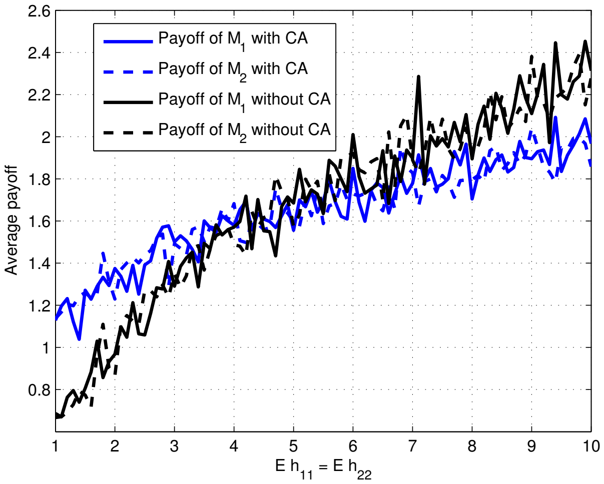


Figure 6.1: Average payoff with NBS fairness

From Chebyshev's inequality, we have

$$\Pr(|\eta_i - \mathbb{E}(\eta_i)| > \chi\sqrt{var(\eta_i)}) \leq \frac{1}{\chi^2}. \tag{6.21}$$

Substituting $\chi = (B_{i-i}R_{ii}[m+1] - \mathbb{E}(\eta_i))/\sqrt{var(\eta_i)}$ and using Eq. (6.20), we can obtain the results in Theorem 1.

■

The above result shows that if the difference between $B_{i-i}R_{ii}[m]$ and $\mathbb{E}(\eta_i)$ is much larger that the variance of η_i, Alg. 7 will converge to the NBE, i.e. $\epsilon \to 0$ when $|B_{i-i}R_{ii}[m]-\mathbb{E}(\eta_i)| \gg var(\eta_i)\mathbb{E}\Delta\varpi_i$. In other words, Alg. 7 will be more useful when the average performance gain brought by DI-CA is large.

6.4 Discussion and Numerical Results

In this section we present numerical results to verify the performance of the spectrum scheduling schemes of DI-CA. We assume the payoff of each MNA is its average downlink channel capacity described in Sec. 6.1 and all the channel gains follow a Rayleigh distribution. In Fig. 6.1, we fix the average value of the channel gains experienced by MNAs in the aggregatable spectrum of each other and consider their performance when the channel gains in their own spectrum change.

It is observed that DI-CA cannot always provide performance improvement over the non-CA case. This verifies the observation that DI-CA can only increase the capacity of each MNO if the capacity obtained from the other MNO's spectrum is higher than that obtained in its own spectrum. It is observed that DI-CA can only provide payoff improvement when the MNAs can obtain a higher capacity sum by sending signals in the aggregatable spectrum of each other. An interesting observation is that the high payoff MNA may switch between the two MNAs when the channel gains change. This is because the NBS fairness always forces at least one MNA to allocate all the aggregatable time portion to the other in exchange for a higher chance of using the aggregatable spectrum. Hence the variance of the channel gains may cause the high payoff MNA to change between M_1 and M_2. Similar observations can be found in Fig. 6.2, where we present the values of p_1^{NBS} and p_2^{NBS} under different channel gains.

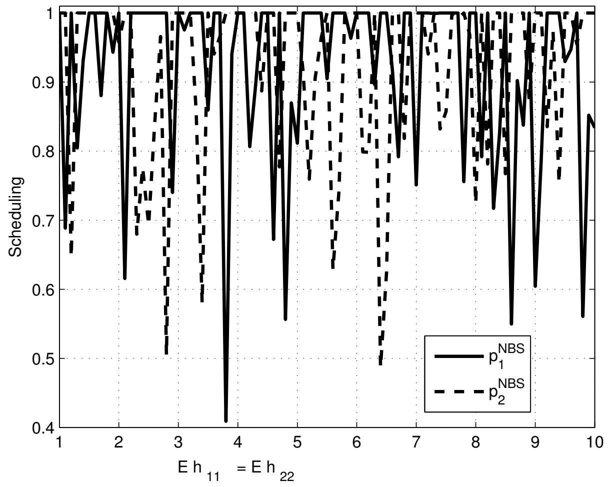


Figure 6.2: Scheduling with NBS fairness

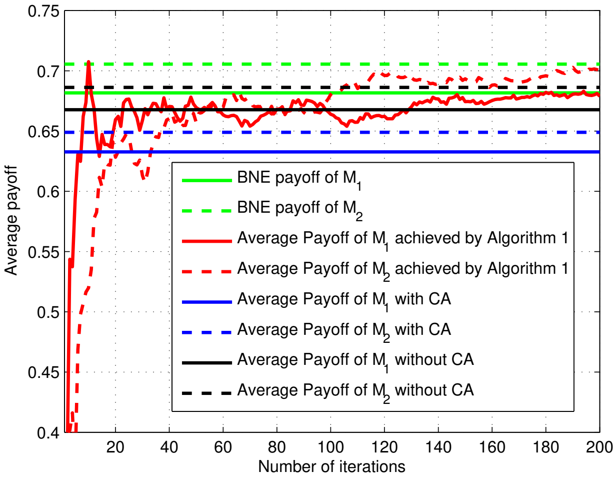


Figure 6.3: Convergence rate of Alg. 7

The convergence of Alg. 7 is illustrated in Fig. 6.3. As Fig. 6.3 clearly shows, the average payoff achieved by Alg. 7 may fluctuate when the number of iterations is small. However, as the number of iteration increases, the performance of Alg. 7 will be higher than either DI-CA or without DI-CA and will eventually approach a BNE.

6.5 Conclusion

This final chapter introduced a DI-CA model to investigate the dynamic scheduling of spectrum resources between independent cellular networks. We derived the conditions under which DI-CA can improve the performance for all the MNAs. We proposed an NBS-based fair spectrum scheduling scheme. To avoid using DI-CA when it cannot provide any performance gain, we devised a Bayesian game-based framework in which each independent operator can decide whether or not DI-CA can improve its performance without knowing the information of others. Finally, we proposed a distributed algorithm to approach a neighborhood of the BNE.

7 Conclusion

THIS thesis analyzed how resource sharing affects different aspects of mobile cellular networks, with a specific focus on network planning decisions. In the following, we first recap the contributions of each chapter and we then propose directions for future research.

7.1 Contributions

In this section we summarize the findings of this thesis dividing them into two categories: (i) infrastructure sharing and (ii) spectrum sharing.

7.1.1 Infrastructure Sharing

The first part of this thesis concentrated on the impact of infrastructure sharing on different aspects of cellular network design, such as network consolidation, network evolution, and how emerging business models can affect deployment decisions of service providers and mobile network operators. Exceptionally we had access to two large-scale datasets. Specifically we had available detailed traffic demand information from two nationwide Irish mobile operators collected at their core networks. Hence, for the first time, we had the chance to compare the actual traffic demand from two operators covering the same nationwide territory. We leveraged this data to study mobile networks in a completely new way by taking into account the actual correlation in space and time of the traffic demand. In this regard, Chapter 3 first presented a quantitative study of traffic demand correlation between two operators, using real traffic and deployment data.

As discussed in Chapter 3, Mobile Network Operators experience low enough correlation, both in space and time, of the traffic demand at their radio access network (RAT). Given that, resource sharing of any form can significantly improve network efficiency. Network topologies can be studied using freely accessible data, and large-scale networks can be studied with a reasonable level of accuracy. We have made available example datasets for other researchers to carry out studies similar to ours at `http://bit.ly/1Fke5xV`.

Infrastructure sharing plays an important role in the network consolidation process and the criteria employed to make consolidation decisions has to consider not only the traffic served but also the trade-offs between savings and quality. In Chapter 4 we discussed a greedy approach to perform network consolidation which results in close to optimal solutions. The analysis used a generalized framework based on a combination of demographic information, network topology, and traffic demand data. Within this framework, we proceeded in Chapter 4 to model the process of modernization of mobile networks as a series of optimization problems; we have proposed different solution strategies and several algorithms to optimize when and where network upgrades and decommissions should take place in a cost effective manner.

The discussion made clear that the evolution of a mobile network is heavily affected by infrastructure sharing agreements and competition regulations. In fact we noted the ability to share infrastructure essentially moves capacity from rural, sparsely populated areas (where some of the existing infrastructures can be decommissioned) to urban ones (where most of the next-generation base stations would be deployed), with a significant increase in resource efficiency. As discussed in Chapter 4, tight competition regulation limits to some extent the ability to share but does not entirely jeopardize those gains, while having the secondary effect of encouraging (or perhaps, forcing) a wider deployment of next-generation technologies.

Emerging business models can shape the way mobile networks will look. As such, we explore the factors that will impact *service-driven networks* as a result of the introduction of new technologies, regulatory decisions, advancements in network virtualizations and the spatial characteristics of the demand. In Chapter 5 we found that the factors with the deepest influence on the final network are not technical but rather economic (e.g., the cost of base stations) and regulatory (e.g., the type of spectrum and technologies OTTs are allowed to use). Different combinations of these factors yield radically different networks.

7.1.2 Spectrum Sharing

We have also analyzed how spectrum sharing (in the form of spectrum/carrier aggregation) affects the decisions of a mobile operator to let a competing mobile operator access its spectral resources. This study was directed towards modelling rational decisions of mobile operators, and to this end, a game theoretic framework has been employed. The contributions of this study are presented and discussed in Chapter 6 of this thesis. There we first derived the conditions under which spectrum aggregation can improve the performance for all the mobile network operators. It was made clear that dynamic spectrum aggregation can take place only if these conditions are satisfied. We proposed a Nash Bargain Solution (NBS)-based fair spectrum scheduling scheme. To avoid using dynamic inter-operator carrier aggregation (DI-CA) when it cannot provide any performance gain, we devised a Bayesian game-based framework. In this framework each independent operator can decide whether or not DI-CA can improve its performance without knowing the information of others and we designed a distributed algorithm to approach a neighborhood of the Bayesian Nash Equilibrium. The advantage of this approach is that it does not require high overhead in the amount of information exchanged.

7.1.3 Summary of findings

The findings have a number of important implications. One of them is of a methodological nature and it is possible to study network sharing and planning decisions using real data with limited extra efforts. From our analysis, it clearly emerged that high gains can be achieved by merging two well established networks and that mobile operators should also consider new network business models that include sharing. However, it is important to identify whether competition levels are high enough in the market, acknowledging that competition should be understood and interpreted differently in rural areas and urban ones. As such, regulators may have to intervene by, for example, enforcing the deployment of extra capacity to be made accessible by new entrants.

7.2 Future Works

The work conducted in this thesis raises some intriguing issues which merit further investigation. We now discuss future research directions dividing them broadly into cost modelling, trade-offs between spectrum sharing and infrastructure sharing and service-drive network planning.

7.2.1 Cost Modelling

To start with, in Chapter 4 we considered the impact of infrastructure sharing on the consolidation and the evolution of mobile networks. To this end, we have assumed that base stations can be used as proxy for evaluating the cost of running a network. The overall process of network planning takes into account a complex mix of technical and economic factors, from OPEX/CAPEX associated with different radio access technologies to possible market advantages over one's competitors. Hence, one avenue of future research would be to incorporate into our optimization analysis a much more sophisticated and realistic cost model. A more elaborate cost modelling clearly requires information that is not publicly available. In addition, the cost figures available for such endeavor are rather inconsistent, as we have already discussed in Chapter 5 of this thesis. In this regard, interviews with mobile operators and equipment vendors are highly desirable in order to obtain more detailed cost in obtaining useful information. This strategy has been implemented in [58] which showed a particular cost structure with reference to the Swedish market.

7.2.2 Trade-offs between Spectrum Sharing and Infrastructure Sharing

Throughout this thesis we have treated infrastructure sharing and spectrum sharing as separate strategies that operators can exploit for their own advantage. However, spectrum and infrastructure sharing may also be applied jointly. In such situations full benefits of network sharing can be achieved when virtualized access networks are set up from a pool of virtualized physical resources including base stations, spectrum or cloud processing units [8]. To this end, we have started exploring the fundamental trade-offs between spectrum and infrastructure sharing at various degrees of spatial correlation between the networks of sharing operators. In addition we started analyzing whether or not as well as to what extent, infrastructure sharing can be substituted for spectrum sharing [129].

We noted that each type of sharing has its own distinctive trait, which we summarize in Fig. 7.1. With reference to Fig. 7.1, each square corresponds to the network performance of a different combination of spectrum and infrastructure sharing. Each square is split evenly between network coverage and user data rate. Saturation of the green-coloured cells reflects the gain with respect to the no sharing case (grey cells), while red-coloured cells represent negative gain. Infrastructure sharing in isolation greatly improves network coverage, but provides minor gain in terms of data rate; spectrum sharing (when worst case interference scenario is considered) provides minor gains in data rate, but degrades coverage. When applied in combination, both coverage and data rate are improved over the no sharing case, with network coverage being slightly worse than in the pure infrastructure sharing case, as a result of increased interference. Hence, infrastructure and spectrum sharing cannot be simply substituted for each other, as there exists a trade-off in the coverage and data rate performance between the two.

We performed our initial analysis by applying stochastic geometry. The key motivation is to analyze the network performance over various realizations of a network consisting of base stations which follow a particular spatial structure. This structure may be expressed as a point process so that the spatial distribution of nodes resembles that of a real radio access network. In this way we

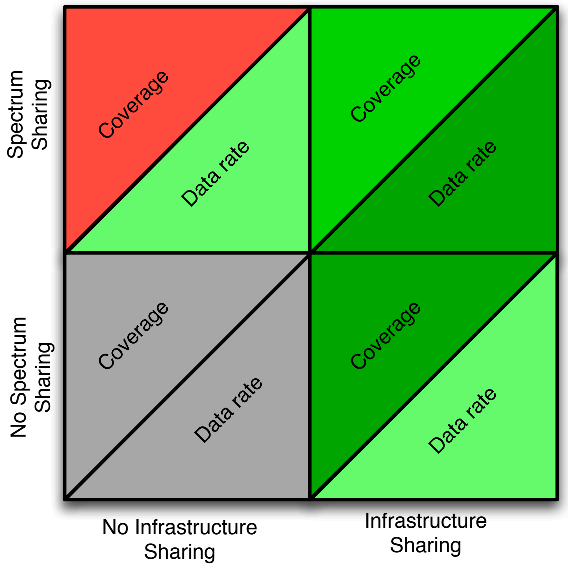


Figure 7.1: Conceptual depiction of the effect of radio access infrastructure and spectrum sharing on network coverage and data rate.

were able to derive in some cases closed form results, while when closed form derivations were not possible, Monte-Carlo simulations were used. Results for specific case studies using real base station deployments, e.g., Ireland, Poland, and UK can be easily obtained. From the analysis it emerged that when mobile operators deployments are highly correlated as it is the case for real-deployments, any gains attainable from aggregating spectrum are significantly reduced. This is because of the increase in interference due to the assumption that base stations operating in the same bands always interfere. This finding stresses the importance of smart (both centralized and distributed) resource management (e.g., scheduling, channel assignment, etc.) mechanisms to counter-balance the potential increase in interference due to spectrum sharing. An immediate extension of this work involves studying the impact of coordinated spectrum sharing techniques on the performance and trade-offs observed.

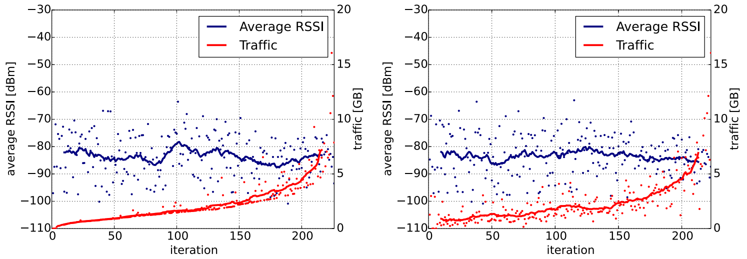
A parallel line of research can potentially involve a broader, multi-disciplinary study that considers the relationship between the economic and regulatory aspects of various forms of wireless network sharing, such as costs and revenues, and the spatial distribution of a shared network.

7.2.3 Service-Driven Network Planning

Promising extensions can be also applied to Chapter 5. The results showing the extension to the case where the MNO stipulates SLA with multiple OTTs can be obtained trivially and our proposed generalized model already addresses this case. However, less intuitive is the case where OTTs can decide to stipulate a contract among multiple MNOs. Our current model does not include competition among multiple MNOs. A single MNO on a market is much more likely to set up higher prices than if there were multiple MNOs present. For example, we could model the interaction among MNOs with a game theory-based framework.

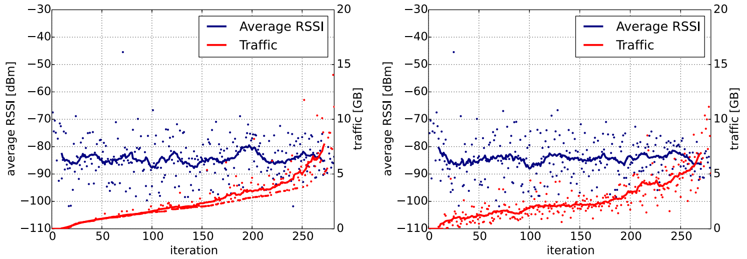
Appendices

A Algorithm Effect on Single Operators



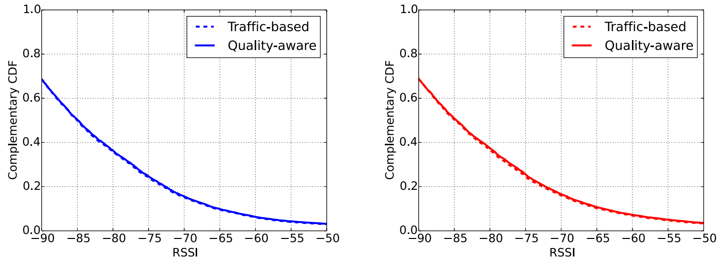
(a) (b)

Figure A.1: MNO$_1$ case. Traffic and average RSSI from covered subscriber clusters of the base station decommissioned at each iteration under the traffic-based (a) and quality-aware (b) usefulness metrics. Dots correspond to individual iterations; lines show the moving average.



(a) (b)

Figure A.2: MNO$_2$ case. Traffic and average RSSI from covered subscriber clusters of the base station decommissioned at each iteration under the traffic-based (a) and quality-aware (b) usefulness metrics. Dots correspond to individual iterations; lines show the moving average.



(a) (b)

Figure A.3: Percentage of the traffic served with RSSI for the shared network under the traffic-based and quality-aware metrics. The savings are set to 15%. MNO$_1$ (a), MNO$_2$ (b).

B Demand Complementarity

Demand modelling

Demand and traffic profiles. As a result of the processing procedure applied to our traces and described in Sec. 3.2, we obtain a set of *demand points* $\varphi_{\mathcal{D}}$, which can be considered as a realization of a marked point process consisting of the following pairs (x, d), where elements $x \in \varphi$ denote locations of points and $d \in \mathcal{D}$ denote the amount of demand units assigned to each point. Moreover, φ is a set of unmarked points in W, where $W \subset \mathbb{R}^2$ is our desired region of interest, and $\mathcal{D}$ is the set of possible demand volumes.

Each demand volume consists of a number of data flows coming from various applications. In order to represent these data flows correctly, we specify applications distribution in a given demand volume of a particular cell. For example, assuming there are $\mathcal{N}$ available applications we could assume uniform distribution of applications in any given demand volume. However, as observed in [79], there is no single traffic profile that can describe applications distribution across space. Yet, following the k-clustering approach, a finite number of traffic profiles can be identified. Hence, we define the set of traffic profiles $\mathcal{P}$, which consists of four types of traffic profiles.

Complementarity of Demand

Given the above model, the goal of the following procedure is to associate the traffic profiles with demand locations to create a marked point pattern $\varphi_{\mathcal{P}}$ consisting of the following pairs (x, p), where elements $x \in \varphi$ denote locations of demand points and $p \in \mathcal{P}$ denote the traffic profile associated with a particular location. The association should be made so to meet the pre-specified complementarity level ρ. In the following we will assume that the demand complementarity is an arbitrary weight that drives our procedure towards either more clustered demand distribution, when $\rho \to 1$, or more uniformly distributed demand, when $\rho \to 0$.

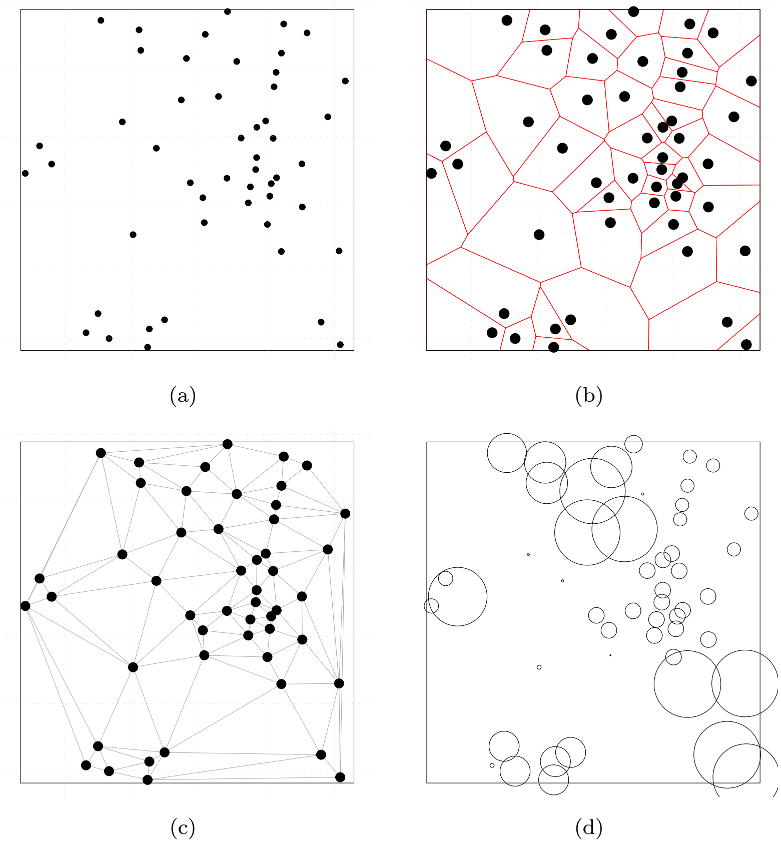


Figure B.1: Converting a point pattern to a graph. We begin by considering the spatial position of each demand point in φ (a). Then, we perform a Voronoi tessellation, where each location is assigned a tile (b). Locations are subsequently associated to vertices of a graph, with edges being drawn between locations whose tiles are neighboring (c). Finally, using the adjacency graph, we spatially allocate traffic profiles to each the demand points (d).

Graph representation of demand. We have already pre-processed our demand data into spatially distributed points. In order to study various complementarity levels of the traffic profiles $\mathcal{P}$, we use graph representation of the demand points, were vertices represent points and edges connect adjacent points. More precisely, as depicted in Fig. B.1, we:

- consider the spatial position of each point (Fig. B.1(a)),
- draw Voronoi tiles [130, Ch. 7] associated with these points (Fig. B.1(b)),
- create an edge between the points that share a common boundary (Fig. B.1(c)).

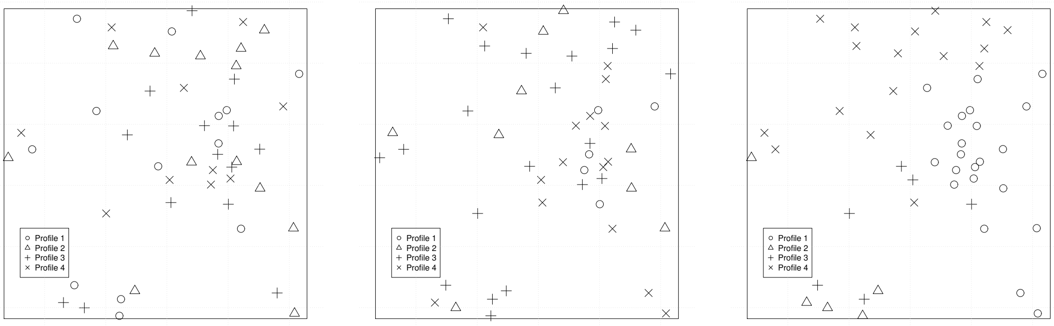
In this way the demand points have been transformed to an adjacency graph G, which allows us to readily apply the following procedure, see Alg. 8, that computes spatial traffic profile allocations for a target complementarity level ρ. What we get as a result of this procedure is a marked point pattern, where points denote fixed demand locations and marks attached to points denote the traffic profiles assigned (Fig. B.1(d)).

Allocation procedure. We initiate the assignment procedure by randomly selecting $|\mathcal{P}|$ points and assigning a unique traffic profile from $\mathcal{P}$ to each of the points. The remaining points form a set of unmarked points φ_0. In each iteration of our procedure we randomly select a point from φ_0 (Line 2) and generate a random weight (Line 3). Next, we compare this weight against our pre-specified complementarity level ρ (Line 4). If the weight is smaller than the target complementarity level, we assign a mark to the point based on the histogram of marks of its adjacent neighbours[1] in G, denoted as $\text{adj}_G(x)$ (Line 5). Else, we assign a random mark generated from the pre-specified mark distribution (Line 7). Eventually, we remove x from φ_0 (Line 8), and we continue the above procedure until all unmarked points are assigned a traffic profile (Line 9).

Algorithm 8 Assigning marks (traffic profiles) to demand points.

```
1: repeat
2:     $x \leftarrow y, y \in \varphi_0$
3:     $w \leftarrow z, z \in \{0..1\}$
4:     if $w < \rho$ then
5:         $m \leftarrow v, v \in \text{adj}_G(x)$
6:     else
7:         $m \leftarrow v, v \in \mathcal{P}$
8:     $\varphi_0 \leftarrow \varphi_0 \setminus \{x\}$
9: until $\varphi_0 = \emptyset$
```

Sample results. In order to initially evaluate the behaviour of our procedure we have looked at the traffic profile assignment for a subset of our data. In Fig. B.2 we can see how the complementarity level affects the distribution of traffic profiles. Clearly, when complementarity is 0 we observe high uniformity, while with growing complementarity we observe an increase in clustering of traffic profiles, i.e., similar traffic profiles group into spatial clusters.



(a) (b) (c)

Figure B.2: Complementarity between different traffic profiles. Each subfigure contains a realization of the traffic profile assignment for a given complementarity level: $\rho = 0.0$ (a), $\rho = 0.5$ (b), $\rho = 1.0$ (d).

Traffic profile vs content distribution. In Table B.1, $f(p)$ - is the spatial frequency, according to uniform distribution, of traffic pattern p, $w(p)$ is the fraction of OTT content in the demand

[1] In the case x has no marked neighbours, we select an unmarked demand point once again.

volume of traffic pattern p, and $w'(p)$ is the fraction of non-OTT content in the demand volume of traffic pattern p. Of course, those percentages apply to the demand levels inside a specific user cluster (point), and every such cluster is assigned one of the four traffic profiles.

In order to calculate calculate the fraction of the OTT content in the total demand, one can apply the following formula:

$$w_{\text{tot}} = \sum_{p \in \mathcal{P}} f(p)w(p) \tag{B.1}$$

where $\mathcal{P}$ is the set of traffic profiles. In the simulations we have used uniform spatial distribution distribution of the traffic patterns, therefore the total fraction of demand in our case equals to 14.75%.

Table B.1: Distribution of the OTT content inside the traffic profiles

Traffic profile	Spatial frequency $f(p)$ (uniform)	Fraction of OTT content $w(p)$	Fraction of non-OTT content $w'(p)$
profile 1	25%	43%	57%
profile 2	25%	10%	90%
profile 3	25%	5%	95%
profile 4	25%	1%	99%

Bibliography

[1] J. Markendahl, “Mobile Network Operators and Cooperation: A Tele-Economic Study of Infrastructure sharing and Mobile Payment Services,” Ph.D. dissertation, KTH, School of Information and Communication Technology (ICT), Communication Systems, CoS, 2011.

[2] L. Cricelli, M. Grimaldi, and N. L. Ghiron, “The Competition among Mobile Network Operators in the Telecommunication Supply Chain,” *International Journal of Production Economics*, vol. 131, no. 1, pp. 22–29, 2011.

[3] B. S. Arnaud, “iPhone slowing down the Internet – Desperate need for 5G R&E networks,” 2012. [Online]. Available: http://billstarnaud.blogspot.com/2010/04/iphone-slowing-down-internet-desperate.html

[4] Credit Suisse, “U.S. wireless networks running at 80% of capacity,” 2011. [Online]. Available: http://benton.org/node/81874

[5] 3GPP, “Network Sharing; Architecture and Functional Description,” 3rd Generation Partnership Project (3GPP), TS 23.251, 2007.

[6] D. E. Meddour, T. Rasheed, and Y. Gourhant, “On the Role of Instrastructure Sharing for Mobile Network Operators in Emerging Markets,” *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 55, no. 7, 2011.

[7] T. Frisanco, P. Tafertshofer, P. Lurin, and R. Ang, “Infrastructure Sharing and Shared Operations for Mobile Network Operator,” in *IEEE Network Operations and Management Symposium*, 2008.

[8] L. Doyle, J. Kibiłda, T. Forde, and L. A. DaSilva, “Spectrum without Bounds, Networks without Borders,” *Proceedings of the IEEE*, vol. 102, no. 3, pp. 351–365, 2014.

[9] M. Balon and B. Liau, “Mobile Virtual Network Operator - Architectural Evolution and Economic Stakes,” in *Telecommunications Network Strategy and Planning Symposium (NETWORKS)*, 2012.

[10] L. A. DaSilva, J. Kibiłda, P. Di Francesco, T. Forde, and L. Doyle, “Customized Services over Virtual Wireless Networks: The Path towards Networks without Borders,” in *Future Network and Mobile Summit*, 2013.

[11] C. Liang and F. R. Yu, “Wireless Network Virtualization: A Survey, Some Research Issues and Challenges,” *IEEE Communications Surveys and Tutorials*, 2014.

[12] 3GPP, “Service Aspects and Requirements for Network Sharing,” 3rd Generation Partnership Project (3GPP), TR 22.951, 2007. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/22951.htm

[13] 3GPP, “Technical Specification Group Services and System Aspects; Study on Radio Access Network (RAN) Sharing Enhancements,” 3rd Generation Partnership Project (3GPP), TS 22.852, 2014.

[14] Radio Access Network sharing agreement between Telia Denmark A/S and Telenor A/S. [Online]. Available: http://en.kfst.dk/Indhold-KFST/English/Decisions/20120229-Radio-Access-Network-sharing-agreement-between-Telia-Denmark-and-Telenor?tc=F1618B4A5A8D4BBC9414DAA8FC6CD519

[15] J. Markendahl and B. G. Molleryd, "On co-opetition between mobile network operators: Why and how competitors cooperate," International Telecommunications Society (ITS), 19th ITS Biennial Conference, Bangkok 2012: Moving Forward with Future Technologies - Opening a Platform for All 72491, 2012.

[16] M. Choudhary, H. Babar, H. Shakeel, and A. Abbas, "Economics of network sharing - a case study of mobile telecom sector in pakistan," in *Collaborative Computing: Networking, Applications and Worksharing, 2009. CollaborateCom 2009. 5th International Conference on*, Washigton DC, Nov. 2009.

[17] J. S. Panchal, R. D. Yates, and M. M. Buddhikot, "Mobile Network Resource Sharing Options: Performance Comparisons," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4470–4482, 2013.

[18] S. Hua, P. Liu, and S. S. Panwar, "The Urge to Merge: When Cellular Service Providers Pool Capacity," in *IEEE International Conference on Communications (ICC)*, Ottawa, Jun. 2012.

[19] M. A. Marsan and M. Meo, "Energy efficient management of two cellular access networks," *SIGMETRICS Perform. Eval. Rev.*, 2010. [Online]. Available: http://doi.acm.org/10.1145/1773394.1773406

[20] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 56–61, June 2011.

[21] GSMA, "Mobile Infrastructure Sharing," Tech. Rep., 2012. [Online]. Available: http://www.gsma.com/publicpolicy/wp-content/uploads/2012/09/Mobile-Infrastructure-sharing.pdf

[22] "Strategic Review of Digital Communications," *OFCOM White Paper*, 2015. [Online]. Available: http://stakeholders.ofcom.org.uk/binaries/consultations/dcr_discussion/summary/digital-comms-review.pdf

[23] C. Beckman and G. Smith, "Shared Networks: Making Wireless Communication Affordable," *IEEE Wireless Communications*, 2005.

[24] J. Hutell, K. Johansson, and J. Markendahl, "Business Models and Resource Managment for Shared Wireless Networks," in *IEEE Vehicular Technology Conference*, 2004.

[25] U.S. Department of Justice and the Federal Trade Commission, "Horizontal merger guidelines," Tech. Rep., Aug. 2010.

[26] European Commission, "EU Competition Law Rules Applicable to Merger Control Situation as at 1 April 2010," Tech. Rep., 2010.

[27] T. Weiss and F. Jondral, "Spectrum Pooling: an Innovative Strategy for the Enhancement of Spectrum Efficiency," *IEEE Communications Magazine*, vol. 42, no. 3, 2004.

[28] M. Buddhikot, P. Kolodzy, S. Miller, K. Ryan, and J. Evans, "DIMSUM-net: New Directions in Wireless Networking Using Coordinated Dynamic Spectrum Access," in *IEEE World of Wireless Mobile and Miltimedia Networks (WoWMoM)*, 2005.

[29] European Commission, "https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/com-ssa.pdf," 2012.

[30] United States President's Council of Advisors on Science and Technology (PCAST), 2012. [Online]. Available: https://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast_spectrum_report_final_july_20_2012.pdf

[31] "FCC Releases Rules for Innovative Spectrum Sharing in 3.5 GHz Band," Federal Communications Commission (FCC), Federal Communications Commission (FCC), Tech. Rep., 2015.

[32] E. A. Jorswieck, L. Badia, T. Fahldieck, E. Karipidis, and J. Luo, "Spectrum Sharing Improves the Network Efficiency for Cellular Operators," *IEEE Communications Magazine*, 2014.

[33] M. Bennis, , M. Le Treust, M. Debbah, and J. Lilleberg, "Spectrum Sharing Games on the Interference Channel," in *International Conference on Game Theory for Networks (GameNets)*, 2009.

[34] D. H. Kang, K. W. Sung, and J. Zander, "Cooperation and Competition between Wireless Networks in Shared Spectrum," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on*, 2011.

[35] [Online]. Available: http://www.saphyre.eu/

[36] [Online]. Available: http://www.ict-samurai.eu/

[37] G. T. R. 36.912, "Feasibility Study for further Advancement for E-UTRA (LTE-Advanced)," 3GGP, Tech. Rep., 2010.

[38] M. Weiss, P. Krishnamurthy, L. Doyle, and K. Pelechrinis, "When is electromagnetic spectrum fungible?" in *IEEE New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2012.

[39] M. Gomez and M. Weiss, "How do limitations in spectrum fungibility impact spectrum trading?" in *Telecommunications Policy Research Conference*, September 2013. [Online]. Available: http://d-scholarship.pitt.edu/19663/

[40] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio Access Network Virtualization for Future Mobile Carrier Networks," *IEEE Communications Magazine*, vol. 51, no. 7, 2013.

[41] "Google – Project Fi," 2015. [Online]. Available: https://fi.google.com/about/

[42] B. Popper, "http://www.theverge.com/2015/4/22/8471243/google-project-fi-mvno-sprint-t-mobile."

[43] "Fast and free facebook mobile access with 0.facebook.com," 2015. [Online]. Available: https://www.facebook.com/notes/facebook/fast-and-free-facebook-mobile-access-with-0facebookcom/391295167130

[44] S. Datoo, "Twitter's latest deal points to ambitions in emerging markets," *The Guardian*, 2013.

[45] M. Nawrocki, H. Aghvami, and M. Dohler, *Understanding UMTS Radio Network Modelling, Planning and Automated Optimisation: Theory and Practice.* John Wiley & Sons, 2006.

[46] E. Amaldi, A. C. abd F. Malucelli, and C. Mannino, "Optimization Problems and Models for Planning Cellular Networks," in *Handbook of optimization in telecommunications.* Springer, 2006.

[47] E. Amaldi, A. Capone, and F. Malucelli, "Planning UMTS Base Station Location: Optimization Models with Power Control and Algorithms," *IEEE Transactions on Wireless Communications*, vol. 2, no. 5, pp. 939–952, 2003.

[48] C. Lee and H. Kang, "Cell Planning with Capacity Expansion in Mobile Communications: a TABU Search Approach," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 5, 2000.

[49] A. Abdel-Khalek, L. Al-Janj, and Z. Dawy, "Optimization Models and Algorithms for Joint Uplink/Downlink UMTS Radio Network Planning with SIR-Based Power Control," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1612–1625, 2011.

[50] L. Shangyun and M. St-Hilaire, "A Genetic Algorithm for the Global Planning Problem of UMTS Netowrks," in *IEEE Global Communications Conference (GLOBECOM)*, Miami, Dec. 2010.

[51] F. Gordejuela-Sanchez and J. Zhang, "LTE Access Network Planning adn Optimization: A Service Oriented and Technology-Specific Prospective," in *IEEE Global Communications Conference (GLOBECOM)*, Honolulu, Dec. 2009.

[52] A. Guo and M. Haenggi, "Spatial Stochastic Models and Metrics for the Structure of Base Stations in Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5800–5812, 2013.

[53] H. Ghazzai, E. Yaacoub, M. S. Alouini, Z. Dawy, and A. Abu-Dayya, "Optimized LTE Cell Planning with Varying Spatial and Temporal User Densities," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, 2015.

[54] S. Boiardi, A. Capone, and B. Sanso, "Radio Planning of Energy-Efficient Cellular Netowrks," in *Internation Conference on Computer Communications and Networks (ICCCN)*, 2012.

[55] D. Amzallag, R. Engelberg, J. Naor, and D. Raz, "Capacitated Cell Planning of 4G Cellular Networks," Computer Science Department Technion, Israel, Tech. Rep., 2008.

[56] J. Kibiłda and L. A. DaSilva, "Efficient Coverage through Inter-operator Infrastructure Sharing in Mobile Networks," in *Wireless Days*, 2013.

[57] L. Cano, A. Capone, G. Carello, and M. Cesana, "Evaluating the Performance of Infrastructure Sharing in Mobile Radio Networks," in *IEEE International Conference on Communications (ICC)*, 2015.

[58] K. Johansson, "Cost Effective Deployment Strategies for Heterogeneous Wireless Networks," Ph.D. dissertation, KTH, School of Information and Communication Technology (ICT), Communication Systems, CoS, 2007.

[59] M. Paolini, "The Economics of Small Cells and Wi-Fi offload," Senza Fili Consulting, Tech. Rep., 2012.

[60] E. Hossain, "Evolution Toward 5G Cellular Networks: A Radio Resource and Interference Managment Perspective," in *IEEE International Conference on Communications (ICC) Tutorials*, 2014.

[61] B. Mölleryd and J. Zander, "Valuation of Spectrum for Mobile Broadband Services - Engineering Value versus Willingness to Pay," in *22^nd European Regional Conference of the International Telecommunications Society (ITS)*, 2011.

[62] S. Han, K. W. Sung, and J. Zander, "An Economic Cost Model for Network Deployment and Spectrum in Wireless Networks," in *24^th European Regional Conference of the International Telecommunications Society (ITS)*, 2013.

[63] M. Michalopoulou, J. Riihijärvi, and P. Mähönen, "Studying the Relationships between Spatial Structures of Wireless Networks and Population Densities," in *IEEE Global Communications Conference (GLOBECOM)*, 2010.

[64] K. Tutschku and P. Tran-Gia, "Spatial Traffic Estimation and Characterization for Mobile Communication Network Design," *IEEE Journal on Selected Areas in Communications*, 1998.

[65] "3GPP TR 36.814: Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," 2010.

[66] J. Riihijärvi, M. Petrova, and P. Mähönen, "Influence of Node Location Distributions on the Structure of Ad Hoc and Mesh Networks," in *IEEE Global Communications Conference (GLOBECOM)*, 2008.

[67] [Online]. Available: http://www.askcomreg.ie/mobile/siteviewer.273.LE.asp

[68] [Online]. Available: http://sitefinder.ofcom.org.uk

[69] [Online]. Available: http://www.arpa.emr.it/pubblicazioni/cem/generale_829.asp

[70] [Online]. Available: http://www.uke.gov.pl

[71] J. Kibiłda, B. Galkin, and L. A. DaSilva, "Modelling Multi-operator Base Station Deployment Patterns in Cellular Networks," *IEEE Transaction on Mobile Computing, accepted with revisions.*

[72] S. Yin, D. Chen, Q. Zhang, and M. Liu, "Mining Spectrum Usage Data: a Large-scale Spectrum Measurement Study," *IEEE Transactions on Mobile Computing*, vol. 11, no. 6, pp. 1033–1046, 2012.

[73] O. Holland, P. Cordier, M. Muck, L. Mazet, C. Klock, and T. Renk, "Spectrum Power Measurements in 2G and 3G Cellular Phone Bands during the 2006 Football World Cup in Germany," in *IEEE New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2007.

[74] T. Kamakaris, M. Buddhikot, and R. Iyer, "A Case for Coordinated Dynamic Spectrum Access in Cellular Networks," in *IEEE New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2005.

[75] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary Users in Cellular Networks: A Large-Scale Measurement Study," in *IEEE New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2008.

[76] R. Keralapura, A. Nucci, Z. L. Zhang, and L. Gao, "Profiling users in a 3G network using hourglass co-clustering," in *ACM International Conference on Mobile Computing and Networking (MobiCom)*, Las Vegas, Sep. 2011.

[77] 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 Conference*, 2009.

[78] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying Diverse Usage Behaviors of Smartphone Apps," in *Proceedings of the 11th ACM SIGCOMM conference on Internet measurement conference*, 2011.

[79] M. Shafiq, L. Ji, A. Liu, J. Pang, and J. Wang, "Geospatial and Temporal Dynamics of Application Usage in Cellular Data Networks," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–1, September 2014.

[80] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding Traffic Dynamics in Cellular Data Networks," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2011.

[81] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Netowrk," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2012.

[82] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," in *ACM SIGMETRICS*, 2011.

[83] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1-2, pp. 17–23, 1950.

[84] A. Palaios, J. Riihijärvi, O. Holland, A. Achtzehn, and P. Mähönen, "Measurements of Spectrum Use in London: Exploratory Data Analysis and Study of Temporal, Spatial and Frequency-Domain Dynamics," in *IEEE New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2012.

[85] L. Anselin, "Local indicators of spatial association – LISA," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[86] F. Malandrino, M. Kurant, A. Markopoulou, C. Westphal, and U. C. Kozat, "Proactive seeding for information cascades in cellular networks," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2012.

[87] F. Pozzi and C. Small, "Analysis of urban land cover and population density in the United States," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 6, 2005.

[88] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Transactions on Vehicular Technology*, 1980.

[89] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial Modeling of the Traffic Density in Cellular Networks," *IEEE Wireless Communications*, 2014.

[90] M. Michalopoulou, J. Riihijärvi, and P. Mähönen, "Towards Characterizing Primary Usage in Cellular Networks: A Traffic-bases Study," in *IEEE New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2011.

[91] RTE News, "Vodafone criticises Three-O2 merger terms," 2014. [Online]. Available: http://www.rte.ie/news/business/2014/0529/620446-vodafone-three-criticism/

[92] European Commission, "REGULATION (EC) No 139/2004 MERGER PROCEDURE," 2012. [Online]. Available: http://ec.europa.eu/competition/mergers/cases/decisions/m6497_20121212_20600_3210969_EN.pdf

[93] 3GPP, "Technical Specification for Physical Layer Aspects of UTRA High Speed Downlink Packet Access, 3G," 3rd Generation Partnership Project (3GPP), TR 25.848, 2007.

[94] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018." [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf

[95] N. Vallina-Rodriguez, V. Erramilli, Y. Grunenberger, L. Gyarmati, N. Laoutaris, R. Stanojevic, and K. Papagiannaki, "When David Helps Goliath: The Case for 3G Onloading," in *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, ser. HotNets-XI. New York, NY, USA: ACM, 2012, pp. 85–90. [Online]. Available: http://doi.acm.org/10.1145/2390231.2390246

[96] B. Leng, P. Mansourifard, and B. Krishnamachari, "Microeconomic Analysis of Base-station Sharing in Green Cellular Networks," in *IEEE International Conference on Computer Communications (INFOCOM)*, Toronto, Apr. 2014.

[97] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3G cellular networks," in *ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2011.

[98] N. Alon, D. Moshkovitz, and S. Safra, "Algorithmic construction of sets for k-restrictions," *ACM Trans. Algorithms*, vol. 2, no. 2, pp. 153–177, Apr. 2006. [Online]. Available: http://doi.acm.org/10.1145/1150334.1150336

[99] Irish CSO census data. [Online]. Available: http://www.cso.ie/en/census/census2011_boundaryfiles/

[100] "LTE Technical Modelling Revised Methodology," *OFCOM White Paper*, 2012. [Online]. Available: http://stakeholders.ofcom.org.uk/binaries/consultations/award-800mhz/annexes/annex14.pdf

[101] R. Xia, M. Rost, and L. E. Holmquist, "Business models in the mobile ecosystem," in *IEEE Conference on Mobile Business and 2010 Ninth Global Mobility Roundtable (ICMB-GMR)*, 2010.

[102] Value Partners, "Mobile 2015: New Spectrum, Different Business Models, More Competition?" [Online]. Available: http://www.valuepartners.com/downloads/PDF_Comunicati/value-partners-120210-mobile-2015-henry-alty.pdf

[103] T. Giles, J. Markendahl, J. Zander, P. Zetterberg, P. Karlsson, and G. Malmgren, "Cost drivers and deployment scenarios for future broadband wireless networks - key research problems and directions for research," in *IEEE Vehicular Technology Conference (VTC) Spring*, 2004.

[104] Facebook Inc. [Online]. Available: http://www.internet.org/

[105] E. Wyatt and N. Cohen, "Comcast and Netflix reach deal on service," *New York Times*, 2014.

[106] H. Zhang, X. Chu, W. Guo, and S. Wang, "Coexistence of wi-fi and heterogeneous small cell networks sharing unlicensed spectrum," *IEEE Communications Magazine*, 2014.

[107] Gurobi Optimization, "http://www.gurobi.com."

[108] A. Caprara, P. Toth, and M. Fischetti, "Algorithms for the set covering problem," *Annals of Operations Research*, vol. 98, no. 1-4, pp. 353–371, 2000.

[109] F. C. Gomes, C. N. Meneses, P. M. Pardalos, and G. V. R. Viana, "Experimental analysis of approximation algorithms for the vertex cover and set covering problems," *Computers & Operations Research*, vol. 33, no. 12, pp. 3520–3534, Dec. 2006.

[110] R. P. Brent, *Algorithms for minimization without derivatives.* Englewood Cliffs, N.J: Prentice-Hall, 1972.

[111] Google Inc., "http://www.google.com/loon/."

[112] Facebook Inc., "https://internet.org/projects."

[113] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimiter Wave Cellular Wireless Networks: Potentials and Challenges," *Proceedings of the IEEE*, 2014.

[114] T. S. Rappaport, S. Randan, and E. Erkip, "Millimeter-Wave Urban Channels Communications: Channel Models, Capacity Limits, Challenges and Opportunities," 2014.

[115] E. Perahia and R. Stacey, *Next generation wireless LANs : 802.11n and 802.11ac.* Cambridge University Press, 2013.

[116] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can wifi deliver?" *Networking, IEEE/ACM Transactions on*, vol. 21, no. 2, pp. 536–550, April 2013.

[117] I. Macaluso, D. Finn, B. Ozgul, and L. DaSilva, "Complexity of spectrum activity and benefits of reinforcement learning for dynamic channel selection," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 11, pp. 2237–2248, November 2013.

[118] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical Analysis and Modelling of Spatial Point Patterns (Statistics in Practice)*, 1st ed. Wiley-Interscience, 2008.

[119] B. Hidalgo and M. Goodman, "Multivariate or Multivariable Regression?" *American Journal of Public Health*, 2013.

[120] N. R. Draper and H. Smith, *Applied Regression Analysis.* Wiley-Interscience, 1998.

[121] Y. Xiao, T. Forde, I. Macaluso, L. DaSilva, and L. Doyle, "Spatial Spectrum Sharing-Based Carrier Aggregation for Heterogeneous Networks," in *IEEE Global Communication Conference (GLOBECOM)*, 2012.

[122] K. Pedersen, F. Frediriksen, C. Rosa, H. Nguyen, L. Garcia, and Y. Wang, "Carrier Aggregation for LTE-Advanced: Functionality and Performance Aspects," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 89–95, 2011.

[123] *Wireless Communications: Principles and Practice.* Prentice Hall PTR, 2011.

[124] Y. Xiao, G. Bi, D. Niyato, and L. A. DaSilva, "A Hierarchical Game Theoretic Framework for Cognitive Radio Networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 10, 2012.

[125] L. Shapley, *A Value for n-Person Games.* Princeton University Press, 1953.

[126] D. Schmeidler, "The Nucleolus of a Characteristic Function Game," *SIAM Journal of Applied Mathematics*, 1969.

[127] A. Muthoo, *Bargaining Theory with Applications.* Cambridge University Press, 1999.

[128] *Stochastic Approximation and Recursive algorithms and Applications.* Springer Verlag, 2003.

[129] J. Kibiłda, P. Di Francesco, F. Malandrino, and L. A. DaSilva, "Infrastructure and Spectrum Sharing Trade-offs in Mobile Networks," in *IEEE New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2015.

[130] M. De Berg, M. Van Kreveld, M. Overmars, and O. C. Schwarzkopf, *Computational geometry.* Springer, 2000.