The Dissertation Committee for Alexander Zemlianov
certifies that this is the approved version of the following dissertation:

Analysis and Design of Wireless Systems with Interface and Provider Diversity: Competition and Cooperation

Committee:

Gustavo de Veciana, Supervisor

Theodore S. Rappaport

Sanjay Shakkottai

John Hasenbein

Ari Arapostathis

Ross Baldick

Analysis and Design of Wireless Systems with Interface and Provider Diversity: Competition and Cooperation

by

Alexander Zemlianov, M.S. Physics, M.S.E.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2005

Dedicated to my parents,

Alexander Petrovich Zemlianov and Natalia Nikolaevna Zemlianova

Acknowledgments

My very special thanks go to my advisor Gustavo de Veciana. Gustavo is a truly exceptional teacher, researcher and a very dedicated guide to (not only his) students. I consider myself very lucky in that I had an opportunity to collaborate with Gustavo throughout my PhD research, since the main theme and directions of this thesis were inspired through many fruitful discussions that we had since I joined Gustavo's group.

I thank all faculty members within the Department of Electrical Engineering for their constant help and support throughout my PhD studies. In particular, many thanks to professor Ari Arapostathis who is responsible for over fifty percent of my knowledge within the area of applied probability – I have taken more classes with him than with any other professor here in the University of Texas at Austin. I owe my knowledge within the areas of networking and optimization to professors Ross Baldick, Sanjay Shakkottai, Scott Nettles and Alan Bovik – these are excellent teachers and I enjoyed their classes immensely. I am also grateful to professors Jeff Andrews, Robert Heath and all members of the Wireless Networking and Communication Group, who played a significant role in attracting me to the field of wireless networking and helped to further and deepen my knowledge within the field. Thanks also to all members of my Qualifying and PhD defense committee, in particular to professors Theodore Rappaport and John Hasenbein for their useful and insightful comments.

I would also like to thank a number of faculty members outside of the ECE Department for the invaluable knowledge that they have passed to me through their teaching. My thanks go to Dr. Takis Konstantopoulos for introducing me to the fields of Stochastic geometry and Wireless Ad Hoc networks – it is through his support that I have gained the background without which this thesis could not have been conceived. I am also thankful to professors Thaleia Zariphopoulou and Stathis Tompaidis for introducing me to the world of applications of probability to stochastic control and finance.

Finally I thank my family, friends and especially my wife, Yekaterina, for their constant support and care throughout the long years of my schooling.

ALEXANDER ZEMLIANOV

The University of Texas at Austin
August 2005

Analysis and Design of Wireless Systems with Interface and Provider Diversity: Competition and Cooperation

Publication No. _____

Alexander Zemlianov, Ph.D.
The University of Texas at Austin, 2005

Supervisor: Gustavo de Veciana

In this thesis we propose research towards evaluating wireless systems which may be based on multiple providers using different technologies, and in which end-systems can select among multiple wireless interfaces and/or modes of communication. A key element in this context is the typically distributed decision making mechanism and associated criterion used by end nodes to select among multiple interfaces or modes of communication. We propose to investigate this problem from two perspectives. First how such decision-making impacts the ability of providers compete with each other. And second, how one might design such decision making mechanisms along with associated network engineering tools so as to minimize cost and optimize system capacity when providers or end-systems choose to cooperate. Our focus will be to investigate the large-scale system performance. As such we propose to devise simple stochastic geometric models capturing the salient features of such systems, e.g., locations of access points and users, coverage areas, spatial nature of available capacity and modeling of decision-making strategies which are spatially dependent. The research presented in this thesis provides a formal basis along with some of the basic insights underlying the design and evaluation of such large-scale

heterogeneous wireless systems.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

1.1.1 Multi-provider scenarios in future wireless networking

The last few decades have been marked by intense growth of wireless communications industry [1]. One attribute of this growth is the complex and heterogeneous networking landscape, integrating the various technologies that have emerged in recent years. The reason for this diversity is twofold. Firstly, emerging technologies have to coexist with legacy systems. Thus new designs can not immediately replace their less efficient predecessors, due to either investors' uncertainty or reluctance caused by possibly large investments that have already been made to predecessor technologies. For example as of now, Code Division Multiple Access (CDMA)-based schemes are widely admitted to be dominant in the next generation of cellular telephony [2]. CDMA was first introduced [3] in 1991, followed by QualComm's adopting it for cellular IS-95 standard in 1993 [4]. It was, however, not until a few years ago that broad recognition of this technology was achieved. Moreover, the European wireless communications market is still dominated by a Time Division Multiple Access (TDMA)-based GSM technology.

Secondly, different technologies can accomplish different goals, and each technology may be advantageous under particular conditions. For example, networks engineered using a currently existing IEEE 802.11b standard can deliver data rates of up to about 11 Mbs. However, since they operate in an unlicensed (ISM) portion of spectrum, they are required by FCC[1] to obey rigid power ceiling constraints. This stipulates extremely limited coverage of such networks, making them suitable only for small (office or home) networking. By contrast, third generation (3G) wide area networks (WAN) technologies strive to achieve a uniform coverage. As they use a quite expensive and limited portion of licensed spectrum, the achieved data rates are much lower. For example in 1XEV-DO (High Data Rate) systems, the achievable down-link

[1] Federal Communication Commission

rates theoretically go up to at most 2.4 Mbs.

It is unlikely that future generations of wireless networks will exhibit reduced complexity. In fact, a large fraction of community researching the new fourth generation (4G) wireless networks concentrates on providing mechanisms for supporting interoperability, e.g., global roaming across *multiple* wireless networks [5] – for example, from a cellular network to a satellite-based network to a high-bandwidth wireless local area network (WLAN). This confirms that 4G wireless networks will be consisting of multiple entities, where each is designed to achieve a particular engineering goal.

It is already frequently the case that wireless networks employing non-interfering technologies, e.g., operating in different portions of spectrum, co-exist within the same spatial locations. We will refer to networks consisting of such non-interfering entities as *heterogeneous* wireless networks. Future generation networks are likely to include joint design of heterogeneous networks, which exploit strengths of various technologies. For example 3G technology can provide ubiquitous mobile access, but performs much worse inside buildings [2]. But then, 3G service could be augmented in each building by installing a non-interfering WLAN within it.

1.1.2 Role of decision making and diversity in wireless connectivity

Guaranteeing seamless transition among heterogeneous networks requires end-devices with *multiple* interfaces that are able to communicate with different network entities. Examples of such devices already exist today. For example, a recently developed "dual-mode" phone has two interfaces which permit it to realize a call through a wide area cellular network or an IEEE 802.11 WLAN access point (hotspot), see e.g., [6]. Users of such devices are able to choose between the services different providers offer so as to exploit the *diversity* of technology/providers. The mode selection mechanism at the end-device could be performed by the user, by observing the quality of received services, monitoring the signal strength, etc. However, a truly seamless implementation would require mode selection to be controlled by a properly configured software client (which we will refer to as "agent"), that resides within the end-device [7].

Clearly, the way that those agents are configured to perform mode selection will affect the ability of the constituent networks to *compete* for a share of the subscribers, and thus affecting the revenues that corresponding service providers are able to collect. Understanding how the decision-making mechanism influences "competitiveness" of different network entities proves useful since it allows one to evaluate the viability of a particular technology with respect to the others. For example, recently there was quite a lot of qualitative discussion of how WiFi technology is able to compete with 3G in providing a ubiquitous wireless access, see e.g. [8, 9, 10]. However no precise convincing analytic models have been proposed and there is still quite a lot of controversy on the conclusions that different groups have reached on this matter.

The way that the mode selection mechanism is implemented at end devices, clearly, also affects the *performance* of heterogeneous systems. Indeed, it is likely that when a large number of agents decide to choose a particular network, that this network would become congested. We envision network systems with *cooperating* constituent entities that are able to dynamically redistribute load, by instantiating particular rules for mode selection across the end-devices and networks. Note that one way to realize globally optimal mode selection is to impose centralized control on the end-nodes' agents – such implementation is suggested in the context of the so-called "tightly coupled" [11] heterogeneous networks, which we discuss in greater length in Chapter 4. In this case the central controller would likely to have to solve a complex optimization problem that involves a lot of information about users' locations, link qualities, amount of interference, current resource congestion levels, etc. However, such an approach will lead to extra complexity and infrastructure investments and is likely to also not scale well [12]. Thus, instead, an attractive design option is a "loosely coupled" architecture, in which mode selections are realized by the agents independently, based on "local" information about the state of the network[2]. The broad research challenge in this direction is effective design of such agents' decision-making strategies, such that cooperative multi-provider networks achieve substantial performance gains.

Allowing for multiple communication modes in wireless devices will also enable more complex cooperation opportunities among wireless *end nodes*. For example, the end-devices that are able to reliably communicate using a particular mode can organize into ad hoc clusters. Nodes within such ad hoc clusters can cooperate, e.g., by relaying each others' packets to extend the reach of a sparse infrastructure [14, 15, 16]; or participate in joint message encoding/decoding, which would enable simple single antenna devices to achieve performance of complex multi-antenna systems [17, 18]. Designing mode selection mechanisms in this setting is even more complex, since there are many opportunities to explore.

1.2 Research objectives and methodology

1.2.1 Objectives

The objective of this dissertation is to investigate how the mode selection (decision making) mechanisms that are employed within end-devices affect the ability of networking entities in a heterogeneous network to *compete* or *cooperate*. We split this complex objective into a series of smaller tasks. We start by studying competition and cooperation between *providers* in a *multi-provider* network. Then we move to analysis of the benefits that the *end devices* and network systems could reap from cooperation, by studying the so-called *hybrid* networks.

[2]Such an approach would be reminiscent to the way CDMA uplink power control is realized [13].

Specifically, we first focus on a scenario where users may choose among two wireless data access providers: a wireless wide area network (WAN) service provider engineered to achieve uniform spatial coverage, and a hotspot provider, i.e., an aggregator of WLAN access points each with limited coverage, and used as throughput enhancements at local hotspots areas. When decision-making is solely controlled by users of wireless devices, we investigate the impact the decision making criterion has on the *competitiveness* of the two networks. We then investigate if *distributed* decision-making mechanisms could be devised (i.e. imposed on end-devices by network designers) so that cooperative multi-provider wireless networks and end-nodes see substantial overall performance gains.

As mentioned above, allowing for cooperation among end nodes will lead to more communication modalities. As there are many ways in which end devices can cooperate, providing a truly optimal mode selection mechanism that covers all cooperation possibilities is a hard task. Our approach to assessing the value of end-node cooperation is to allow for only one additional mode – namely – *ad hoc relaying*. We study a *hybrid* network model, which consists of "ad hoc" nodes, that can relay information among each other via wireless links, and of infrastructure nodes (base stations) that can communicate with ad hoc nodes in a wireless manner are themselves interconnected via independent high capacity wired or wireless links.

Given that communication in such networks could be realized in a variety of different ways, we will investigate what is the "optimal" way to operate such networks. Optimality can be associated with a number of different criteria, in particular, given a particular traffic pattern, we could be interested in ensuring that nodes experience on average the least possible delay, or in guaranteeing that each node sees the largest possible common throughput. Following [19][20] we associate optimality with the network *capacity*, that is we attempt to find communication paradigms (such as optimal scheduling, routing, and communication mode selection at the end nodes) that guarantee each node with the largest common throughput.

1.2.2 A note on methodology

This thesis will emphasize the role of *spatial dimension* in wireless systems. Spatial relationships are clearly fundamental in designing wireless systems. They play a critical role in determining power link budgets, coverage areas, co-cell interference, and spatial reuse in cellular systems. Interest in ad hoc networking, and approaches to enabling wireless access to the Internet, have placed renewed emphasis on the spatial dimension. For example, research on the "capacity scaling" of ad hoc networks with the number of participating nodes has highlighted its fundamental relationship with spatial packings of concurrent transmissions[19, 21]. Similarly, recent innovations at the physical and MAC layers, e.g., opportunistic beamforming and multiple antenna systems, are based on exploiting temporal and spatial diversity seen at end nodes in a wireless environment, see e.g., [22]–[26].

Explicit modeling of spatial relationships among devices and/or access nodes has historically taken either a deterministic approach, e.g., selecting a particular pattern for the locations of access points or devices, or introduced the full complexity of modeling an environment's propagation characteristics, (see, e.g. [27]). Recently, researchers have proposed an alternative approach, based on stochastic geometric models, which permits capturing some of the richness of wireless environments, e.g., channel diversity, interference, and spatial load fluctuations, without resorting to unwieldy complexity [28]–[31]. Our own experience with this approach strongly suggests it provides a valuable tool to address the many challenges faced in engineering wireless networks [32, 33]. In this thesis we stress the role that spatial analysis of wireless systems can play in driving research towards improving old, and devising new, communication modalities.

1.3 Organization of the dissertation

We split the thesis into two parts. Part I deals with analysis of competition and cooperation in a multi-provider network that consists of a WAN and a set of WLANs. Part II focuses on modeling hybrid wireless networks.

Specifically, in Chapter 2 we will introduce a stochastic-geometric model for the locations of users, WAN and WLAN access points, service zones and describe an approach to modeling the dynamics of users' decisions. We will demonstrate how the dynamics eventually reach a fixed point in the space of configurations for users' choices which we will refer to as an "equilibrium". Then we will show how decision-making mechanisms on end-devices affect the structure such equilibria possess. The structure of such equilibria is used to assess the competitiveness of WAN with respect to the WLANs. In particular, our approach allows one to analyze via simulation the dependence of a provider's competitiveness on a set of a few parameters, such as the aggregate bandwidths available at WAN and WLAN access points, spatial densities of access points and users, service prices as well as the characteristics of traffic patterns that users generate. In an effort to provide analytic tools, in Chapter 3 we analyze an asymptotic scenario where the service area of each WAN AP contains a significantly large number of users and hotspots.

In Chapter 4 we shift our focus from the analysis of competition to the analysis of the benefits that the hotspots and WAN could gain from cooperation and study the so-called "loosely coupled" heterogeneous networks. In such networks both WAN and WLANs operate independently, and thus the decision-making has to be located within the dual-mode devices. We demonstrate that WAN is able to statistically multiplex spatial load fluctuations on hotspots, by removing the extra load from overly congested hotspots. Such load-smoothing capability of the WAN, however, can only be exploited if *congestion*-sensitive decision making mecha-

nisms are imposed on end-devices. Finally, in Chapter 4 we consider the problems of backhaul dimensioning, jointly optimized with congestion-sensitive decision-making. We show that our optimal design enables one to achieve target performance in a system while realizing significant savings in resource costs.

In Chapter 5 we complement the work discussed in Chapter 4 via studying "tightly-coupled" heterogeneous wireless systems. In such systems, agents within end-devices are also provided with additional feedback from the different networking components, which enable agents' decisions to be better tuned towards the needs of the overall network. We formulate and evaluate centralized and distributed decision-making algorithms, akin to load-balancing algorithms in wire-line networks, that enable further performance gains in multi-provider wireless systems. The major contribution of Chapter 5 is formulation of decision-making metric which ties together various congestion signals with the amount of interference at particular spatial locations.

Chapter 6 provides background on hybrid networks and their modeling. We briefly review the related work on the scaling of hybrid networks' *capacity*, in the case when the ad hoc and infrastructure communications share channel resources. Previous work on the capacity scaling of hybrid networks assessed only what was possible for a narrow set of routing and scheduling strategies. Our approach provides a rigorous capacity analysis which is much more systematic and general than previous work.

Chapter 7 summarizes the major contributions of this thesis and briefly outlines the possible directions for future research.

Part I

Multi-provider wireless networks

Chapter 2

Model and analysis

In this chapter we develop a model for a multi-provider network that consists of two entities operating within the same area, and using non-interfering technologies: a wireless wide area network (WAN) and a set of wireless local area networks (WLANs/hotspots). The heterogeneous wireless system which we study is an example of a "hierarchical overlay", a "multi-tier", or an "umbrella" network, with macro- and micro-cell levels corresponding to the WAN and WLANs coverage areas respectively. Efficient design of such networks was addressed in the past in the context of cellular voice applications [34]–[37]. It was noticed that introducing an overlay level has a capability to reduce the number of dropped calls due to handovers. In these studies both hierarchical layers operate over a shared spectral band, thus the efforts targeted ways to efficiently allocate spectrum among layers.

Emerging integrated 3G and WiFi networks [7]–[11] and dual-mode wireless devices [6] have stimulated renewed interest in the problem of heterogeneous network design [42]–[44]. The major application of these networks is providing enhanced data services, which have drastically different quality of service requirements than circuit-switched voice. In addition, different networking components may use orthogonal technologies, e.g. they might utilize non-overlapping portions of spectrum. Thus the design tools for hierarchical networks where the layers share spectrum, developed in [34]–[37] for circuit-switched applications can not be readily adopted in the context of heterogeneous wireless data networks.

2.1 Summary of main features

In our model it is assumed that users have dual mode devices which are able to communicate with both hotspots and the WAN access points and thus, users that are covered by both a hotspot and the WAN can select between the two providers by choosing the corresponding interface/mode. As discussed in Chapter 1, mode selection can be carried out by the user of the device or a properly configured software client. To abstract from these two cases, in what

follows we will use the word "agent" to refer to the entity responsible for decision-making at the end-device.

Our setup incorporates two major components: a geometric model for providers' and users' spatial interactions and a utility-based model for agents' decision-making. In particular, to capture the geometry of the network we use a simple geometric framework akin the stochastic-geometric models first introduced in [28]. The basic idea is to represent the locations of subscribers and access points (APs) as realizations of spatial point processes (e.g. Poisson) and the service zones associated with the access points as functionals of the realizations of these processes. The main advantage of such models is that they allow one to analytically capture the effect of spatial and load variations in the system based on a reduced set of salient parameters.

We assume that agents make greedy decisions in an effort to maximize their associated *utility function*, which captures users' valuation of the available services. Each agent switches to the provider offering the higher utility at random times, where those times might, for example, model the closing and initiation of communication sessions. Our geometric model is such that it imposes agents in close vicinity of each other to choose from similar sets of access points. Thus whenever an agent switches to an access point it affects the congestion level that is seen by all nearby users already connected to this access point. As the perceived quality of service, and hence agents' utilities are likely to depend on resource congestion level, the agents' provider selections may change over time. This leads to *dynamics* for the configuration of agents choices, where decisions of agents within a limited vicinity of each other are inter-dependent. We show that these dynamics eventually converges to a not necessarily unique, but fixed configuration, which we refer to as an *equilibrium*.

In Section 2.6, we study the properties of agents' decisions equilibria. In particular we concentrate on the simple, but quite insightful case where utilities depend only on resource congestion level, and deployed hotspots are identical, in that they are characterized by the same bandwidth, coverage radius, etc. We define a notion of *uniqueness* for equilibria and provide conditions under which the equilibrium is unique. Guaranteeing uniqueness will prove useful in two respects. On the one hand, in Chapter 3 it will allow us to devise, an asymptotic estimate for a fraction of agents that choose either of the providers, and hence to estimate the competitiveness of providers with each other. On the other hand, we will show in Chapter 4 that systems with unique equilibrium possess properties that could be exploited when designing cooperative multi-provider systems under a target performance constraint.

2.2 Relation to existing work

Models that account for interactions between decision-making and systems' spatial properties have been considered in a number of applications spanning various disciplines. Thus, e.g. in

economics, one of the most famous models that explicitly analyzes the interplay of decision-making and spatial relationships across agents is the Hotelling spatial competition model [45, 46]. In its simplest version the model considers uniformly distributed customers on a line segment and two initially arbitrarily located firms within the same segment. Customers buy the same goods from the firm closest to them, thus the amount of revenue a firm is able to generate, given some fixed locations of both firms, is proportional to the portion of the line segment that is served by this firm. It is assumed that at each time slot a single firm changes its location so as to serve a larger portion of the segment, thus a firm's decision-making incorporates the information of spatial location of the competing firm.

Similarly, in statistical mechanics, Ising models [47] are used to describe phase transitions in magnetic materials. The simplest Ising model consists of a number of entities (spins) that are in some arbitrary way distributed on a line. The spins can only be in two states ("up" or "down") which are identified by the spin orientation and they interact with each other and an external magnetic field. The energy associated with this interaction depends on the orientation of all individual spins in the system. One can associate changes of orientation of each individual spin (due to physical forces exerted on it) as the spin's decision-making. Here again, changes in spin orientations are implicitly linked to spatial configuration of the system.

The main contribution of our modeling approach is that it explicitly ties various spatial factors (and associated uncertainty) with agents' decision-making in the context of heterogeneous wireless networks. Our geometric model is quite general in that it allows a lot of flexibility in the way the geometry of the network is defined. For example, the locations of access points and users could be assumed either to be deterministic or random, and the service zones associated with access points could be deterministically given or random and dependent on the locations of access points.

Our choice to represent decision-making via utility functions is motivated by the large variety of models studied in economics and game theory, (see, e.g. [48]). Our analysis is on a par with that for game-theoretic models which take into account the effects of agents learning [49, 50] and attempt to not only identify possible equilibria in the system but to study the dynamics of decision-making in time. The study of the system's convergence to equilibrium is also reminiscent of the analysis of convergence problems in distributed computation theory [51] and stability problems addressed in control theory [52]. It turns out, however, that straightforward application of standard control-theoretic tools to analysis of convergence in our system presents certain challenges, due to the fact that we can not impose restrictions on how fast the changes in the system are allowed to occur. Indeed, our model does not assume that users are cooperative which means any user is permitted to redirect all its flow from one access point to another as opposed to gradually shifting it as would be required by slow adaptation mechanisms. However, some existing game-theoretic work answers (however not fully) some of the

questions we pose, and we further comment on this matter in Section 2.5.

2.3 Spatial model and notation

2.3.1 General formulation

In this section we introduce a geometric model that captures the key features of spatial interactions between users, the WAN and WLANs. The setup has two components: (i) a model for locations of access points (APs) of various types, and (ii) a model for service zones associated with each access point. Here and in what follows by the "service zone" of a WAN or a hotspot AP we mean the set of locations on the plane, that the AP can serve.

We will let locations of subscribers, hotspots and WAN APs be represented by the corresponding realization of three *simple*[1] point processes Π^a, Π^h and Π^w within a bounded spatial region D. We will denote the resulting locations of agents, hotspots and WAN AP respectively via $\pi^a \equiv \{a_i\}_{i=1}^I$, $\pi^h \equiv \{h_k\}_{k=1}^K$ and $\pi^w \equiv \{w_m\}_{m=1}^M$ and assume that these locations are independent of time (i.e. agents and access points do not move). With each AP we associate a static (in time) geometric object with (random) size and shape that might depend on the spatial realization of the APs and agents within D – these geometric objects will serve as first order approximations for the APs' service zones. Below we will use S_k^h and S_m^w to refer to the service zones of a hotspot h_k and w_m respectively.

Note that in practice the particular shape of service zones would depend on the underlying technology. Thus, for example, in an IS-95 system a mobile decides whether it belongs to the service zone of a particular AP by comparing the strength of pilot signals in its vicinity. In our model, for the reasons we will explain later, we will impose that service zones obey a set of restrictions, which are listed below:

Assumption 2.1.

1. *The WAN has full coverage, i.e.* $\bigcup_{m=1}^M S_m^w = D$.
2. *The service zones of distinct WAN (WLAN) APs are disjoint, i.e.* $S_{m_1}^w \cap S_{m_2}^w = \emptyset$ *and* $S_{k_1}^h \cap S_{k_2}^h = \emptyset$, *for* $m_1 \neq m_2$, *and* $k_1 \neq k_2$.
3. *The service zone of any WLAN AP is fully contained within a service zone of some WAN AP, i.e. for all* $k = 1, 2, \ldots, K$, $S_k^h \subset S_m^w$, *for some* m, $1 \leq m \leq M$.

Assumption 2.1.1 is typically true when the WAN utilizes a portion of licensed spectrum and APs and mobiles use large enough powers for their transmissions. Assumption 2.1.2 is valid for scenarios where WLANs are sufficiently spaced or when they operate using orthogonal

[1] The location of each WAN or hotspot AP is not shared by any other AP [53], i.e. points do not overlap.

spectra. It also holds for some current WAN technologies, but would not be true for systems implementing a "soft handoff". For these systems a mobile "on the cell boundary" may be simultaneously served by several WAN APs, however Assumption 2.1.3 can still be accepted as a reasonable, first order approximation. Finally, Assumption 2.1.3 is motivated by the fact that WAN and WLAN technologies operate at significantly different spatial coverage scales. We will also impose the following technical assumption on the realization of agents and hotspots within WAN service zones, which is satisfied in all practical situations:

Assumption 2.2. *For all $m \in \mathbb{N}$, the service zones S_m^w contains an almost surely finite number of agents and hotspots.*

We postulate that the agents covered by the service zone of a WAN and a WLAN AP can connect to either. We let C_m be the subset of S_m^w that includes spatial locations where agents would have the option to choose among a hotspot and WAN AP w_m, i.e.:

$$C_m \triangleq \bigcup_{k \in \mathcal{K}_m} S_k^h,$$

where $\mathcal{K}_m$ denotes the set of indices of hotspots located within the S_m^w. (for notation summary, see Table (2.3.1)) Users which fall in $\bar{C}_m \triangleq S_m^w \setminus C_m$ can not make a choice and will be assumed to automatically connect to WAN AP w_m. By contrast, an agent $a_i \in C_m$ is also covered by some hotspot h_k's service zone and can choose between connecting to *either* h_k or the WAN AP w_m. Note that the set of Assumptions 2.1 effectively constraints the agents covered by the same WLAN to select between this WLAN and *the same* WAN AP[2].

2.3.2 Example of defining service zones

We now give a particular example in which we define service zones so that they obey Assumption 2.1. Figure 2.1 a exhibits a realization for a stochastic geometric model for the two competing wireless access providers: WAN base stations are shown as boxes, with associated coverage areas modeled by a Voronoi[3] tessellation, i.e., each access point is responsible for locations which are closest to it[4]. Thus, the WAN provider's service is available at all spatial locations. By contrast, the second provider's WLAN access points, shown as triangles, have limited coverage areas which are modeled by discs centered at each access point. This captures a technology with a highly constrained transmit power, e.g., IEEE 802.11 access points sharing unlicensed spectrum.

[2] As will be seen later this requirement makes each agent's choice contingent on information available locally at WAN AP w_m.

[3] Voronoi cell of $w_m \in \pi^w$ is the set of all points on the plane that are closer to w_m than to any $w_n \in \pi^w$, $n \neq m$.

[4] In practice, there would be overlap among coverage areas associated with base stations, yet this is a reasonable approximation in the case where relatively high power levels are used, see e.g., [54].

Π^a	Point process modeling agents' locations
Π^h	Point process modeling hotspots' locations
Π^w	Point process modeling WAN AP locations'
π^a, π^h, π^w	Realization of Π^a, Π^h, Π^w
$\pi(A)$	All points of a realization π that fall within the set A
$\lvert\pi(A)\rvert$	Number of points in $\pi(A)$
$\lvert x\rvert$	Length of vector $x \in \mathbb{R}^2$
$B(x,r)$	Disc of radius r centered at $x \in \mathbb{R}^2$
V_m^w	Voronoi cell of WAN AP $w_m \in \pi^w$
V_k^h	Voronoi cell of hotspot AP $h_k \in \pi^h$
$\mathcal{K}_m$	$\{k:\ h_k \in \pi^h(V_m^w)\}$, indices of hotspots located within the Voronoi cell V_m^w
S_k^h	Service zone of hotspot h_k
S_m^w	Service zone of WAN AP w_m
C_m	Subset of S_m^w where agents can make choices
$\bar{C}_m$	$S_m^w \setminus C_m$
M_m^w	Total number of agents in S_m^w
M_k^h	Total number of agents in S_k^h
M_{C_m}	Total number of agents in C_m
$M_{\bar{C}_m}$	Total number of agents in $\bar{C}_m$
$N_m^w(t)$	Total number of agents connected to WAN AP w_m at time t
$N_k^h(t)$	Number of agents connected to hotspot h_k at time t, where k is s.t. $a_i \in S_k^h$
$U_i^w\left(N_m^w(t)\right)$	Utility function of agent $a_i \in S_m^w$, connected to WAN AP w_m at time t
$U_j^h\left(N_k^h(t)\right)$	Utility function of agent $a_j \in S_k^h$ connected to hotspot h_k at time t

Table 2.1: Notation Summary

We formally define the service zones as follows. With each hotspot $h_k \in \pi^h$ we associate a disc $B(h_k,d)$ of radius $d>0$ and centered at h_k. We assume that service from h_k is available only within the disc (see Figure 2.1). In addition, we assume that agents desiring to connect to a hotspot will connect only to the *closest* feasible hotspot. This yields a service zone S_k^h for hotspot AP h_k given by:

$$S_k^h \triangleq V_k^h \cap B(h_k,d)\,.$$

For each WAN AP $w_m \in \pi^w$ we define its service zone, S_m^w, to be its Voronoi cell, V_m^w, augmented by the service zones of the hotspots that have their APs within V_m^w:

$$S_m^w = V_m^w \bigcup \left(\bigcup_{k\in\mathcal{K}_m} S_k^h\right) \setminus \left(\bigcup_{l\in\cup_{n\neq m}\mathcal{K}_n} S_l^h\right),$$

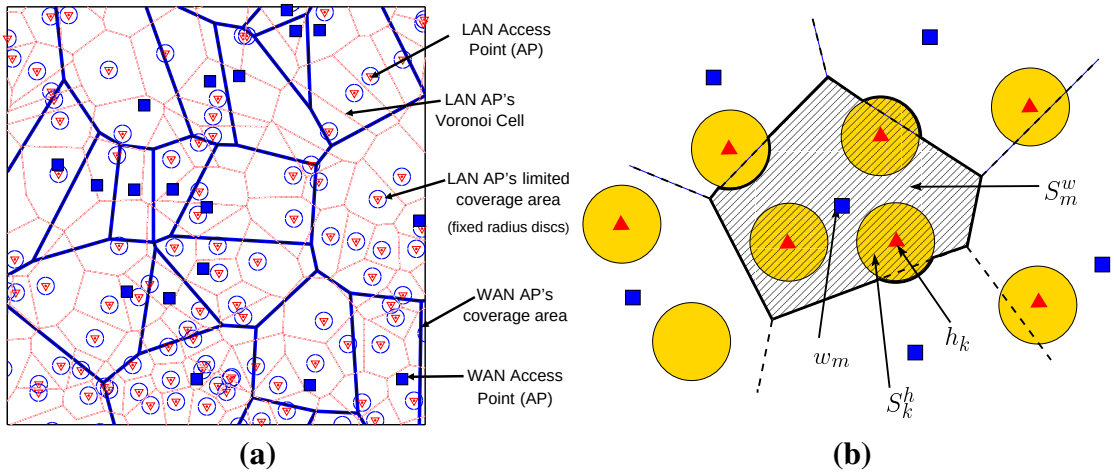


Figure 2.1: Example of defining WAN and WLAN service zones to ensure Assumptions 2.1.1-3. **(a)**: Bird's eye view of network geometry. **(b)**: WAN AP service zone definition.

(in this case $\mathcal{K}_m$ denotes the set of indices of hotspots located within the Voronoi cell V_m^w). Note that each agent $a_i \in \pi^a$ selects between connecting to the closest hotspot AP h_k (if it is covered by its service zone) and the WAN AP w_m *which contains* h_k in its service zone.

2.4 Modeling decision making

Whenever an agent is covered by service zones of two different APs she can potentially choose which AP to connect to. Agents can base these decisions on the distance from an access point, amount of received bandwidth, service price, etc. To make our approach general, we assume that agents base their choices on a particular utility function that captures a user's valuation of the quality of received services. We will consider two types of utility functions which we define below.

Definition 2.1. *We say that the utility function U_j^w (U_j^h) of an agent a_j connected to WAN AP (hotspot AP), is congestion and agent dependent if U_j^w (U_j^h) is a function of a total number of agents connected to the WAN AP (the hotspot AP), and possibly is different for each j.*

Example 2.1. We may have a utility U_j^w for an agent a_j connected to a WAN AP w_m which depends linearly on the throughput T_j^w that the agent obtains by connecting to w_m. Assuming that the WAN provider charges each agent p units of currency per unit service rate that an agent is granted, we have, for some constant $a > 0$:

$$U_j^w = (a - p)T_j^w .$$

Note that T_j^w depends on the resource allocation mechanisms employed at the WAN AP. At the same time, T_j^w may be "agent dependent", in the sense that it may depend on the location of the

agent relative to the WAN AP. For example, WAN APs might serve agents in a time division fashion and grant them equal share of the available time slots. However, the amount of data that the WAN AP is able to deliver to (receive from) agents within their slot shares varies across agents, due to the difference in the quality of their channels.

Let R_j^w denote the data rate of an agent if it was the only one served by the WAN. Clearly R_j^w depends on many factors, e.g. quality of channel available to agent a_j, particular power control strategies used by WAN AP, etc. Thus, for example we might roughly model this as:

$$R_j^w = C\,\mathbb{E}_{\eta_j} \log(1+\eta_j)\,,$$

where $C > 0$ is some constant, and the expectation is taken over the distribution of SINR η_j seen by agent a_j at its location. Now, assuming there are N_m^w agents connected to WAN AP w_m, and equal time allocation among agents, we have $T_j^w = R_j^w / N_m^w$. Hence we arrive at the following expression for utility, which has a congestion and agent dependent structure:

$$U_j^w(N_m^w) = \frac{(a-p)R_j^w}{N_m^w}\,.$$

Example 2.2. Opportunistic scheduling is a technique that is designed to exploit inherent diversity of wireless media and is utilized in many third generation wireless systems, see, e.g. [55]. For example, in Qualcomm's 1XEV-DO systems, a WAN AP serves only one user per time slot with full available power, whereas the selected user is the one achieving the largest current per-slot throughput relative to the average throughput available at the user's location. It can be shown [56], that under the symmetric fading assumption, such scheduling results in a "proportionally fair" throughput gains for each user and thus:

$$T_j^w = \frac{g(N_m^w)R_j^w}{N_m^w}\,,$$

where $g(N)$ is the throughput gain factor which is an increasing function of N. Note that in general $g(N)$ depends not only on the number of served users but also on the channel statistics of all users served by WAN AP w_m. Under the simplifying assumption that $g(N)$ depends solely on the number N of served users[5], one can modify utilities in Example 2.1 to account for opportunism exploited at APs. Then we have for utility function of agent a_j:

$$U_j^w(N_m^w) = (a-p)R_j^w \frac{g(N_m^w)}{N_m^w}\,.$$

[5]This perhaps is not a bad assumption when the number of users is sufficiently large: for Rayleigh fading channel statistics, for example, $g(N)$ can be shown to grow logarithmicaly with N, thus dependence on individual agent's channel is likely to be weak.

Definition 2.2. *We say that a utility function U_j^w (U_j^h) of agent a_j connected to a WAN (hotspot) AP is solely congestion dependent if it depends only on the total number of agents, N, connected to the WAN (hotpsot) AP, and for each $N \in \mathbb{N}$, $U_j^w(N) = U_i^w(N)$ ($U_j^h(N) = U_i^h(N)$) whenever the agents a_j and a_i fall within the service zone of the same WAN (hotspot) AP.*

Example 2.3. In Example 2.2 assume that for all $a_j \in S_m^w$ we have $R_j^w = R_m^w$, for some $R_m^w > 0$, i.e. average over time link data rates are the same for all agents served by the WAN AP w_m. Then we arrive at:

$$U_j^w(N_m^w) = \frac{(a-p)g(N_m^w)}{N_m^w} R_m^w,$$

which has a solely congestion dependent structure.

Example 2.4. Examples 2.1–2.3 consider time-division multiplexed system, where a WAN AP grants equal long-term shares of time slots to all serviced by it users. Thus users with relatively worse channels are at a disadvantage, since they achieve lower throughputs. One can consider a scenario where a WAN AP serves some users longer than others (on average), in order to compensate for this disadvantage.Such service model would be especially appropriate for real-time users, e.g. the ones running (rate-adaptive) multimedia applications. We can find the share of slots that each user is granted, if we assume that the WAN AP provides all users with the largest common long-term throughput. Simple computation shows that the common throughput T seen by all users is given by:

$$T_j^w = T = \frac{1}{\sum_{j=1}^{N_m^w} 1/R_j^w},$$

In the limit of large number of users, T can be approximated as:

$$T = T(N_m^w) \approx \frac{1}{N_m^w \mathbb{E}[1/R_j^w]},$$

where the expectation is taken over the distribution of average over time data rates seen by a typical agent connected to the WAN. Associating utility with the throughput, as before we find:

$$U_j^w(N_m^w) = (1-p)T(N_m^w) = \frac{1-p}{N_m^w \mathbb{E}[1/R_j^w]}.$$

Note that $U_j^w(N_m^w)$ is of solely congestion dependent type.

Example 2.5. Examples 2.1–2.4 are formulated for a scenario in which each user has an infinite backlog of data to download from/send to its WAN AP. In this case users indeed benefit from larger long-term throughputs, and thus using utilities that are functions of the corresponding throughputs is quite natural. In many conventional applications, such as, e.g. web browsing, ftp

transfers, etc., users are interested in downloading finite blocks of data, i.e. files of finite size. In that case a user's appreciation of service at WAN APs will instead be sensitive to the delays that the user experiences during a typical file transfer. One may thus relate utility of an agent to the average of such delay.

Let us assume that all users generate a Poisson stream of γ files per time unit with sizes that follow general distribution (same across all users) with mean f. If we take the assumption of Example 2.3, where the link capacities R_j^w at all spatial locations are the same and equal R^w, and neglect the diversity gain (i.e. $g(N) \equiv 1$), then the average delay for typical transfer is given by that of a single-class processor-sharing queueing system (see, e.g. [57]) thus:

$$\mathbb{E}[D_j^w] = \frac{1}{f/R^w - N_m^w \gamma}.$$

Using this expression, one can construct a solely congestion dependent utility:

$$U_j^w(N_m^w) = -\frac{1}{f/R^w - N_m^w \gamma} - cf,$$

where we took into account that users are penalized by both their transfer delay the cost c that the WAN provider charges them per bit of transfer.

Consider an agent $a_i \in \pi^a(C_m)$ that is connected to WAN AP w_m at time t and assume that the total number of agents that are connected to w_m at that time is $N_m^w(t)$. We model the level of "satisfaction" of agent a_i with the service via a congestion and agent dependent utility function $U_i^w\left(N_m^w(t)\right)$ that depends on the current congestion level and possibly the agent's location[6] within S_m^w.

We assign a solely congestion dependent utility function $U_j^h\left(N_k^h(t)\right)$ to an agent $a_j \in \pi^a(S_k^h)$ connected to a hotspot at time t. Here $N_k^h(t)$ denotes the total number of agents that are connected at time t to the same hotspot as agent a_j. As opposed to the case with service from the WAN, we require that the perception of service from hotspots to be the same for agents connected to the *same* hotspot, i.e., if $a_i, a_j \in S_k^h$, then $U_j^h(N) = U_i^h(N)$, for any $N \in \mathbb{N}$. However, we do not impose this restriction for agents connected to different hotspots, thus we retain the flexibility of including potentially different hotspots' types in the model[7].

In the sequel we will use the following assumption for the utility functions:

Assumption 2.3. *For all $i \in \mathbb{N}$, $U_i^w(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ and $U_i^h(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ are continuous, monotonically decreasing functions.*

[6]Note that this allows to model a situation when the agents, that are farther from the WAN AP have potentially worse communication channels.

[7]For example, hotspots could support different bandwidths.

Once utility functions have been specified for each agent, we will assume agents make decisions consistent with maximizing their utility, i.e., connect to the provider offering the higher utility. However, we will account for a fixed cost of switching to another interface. We stress here that this is simply a model for decision-making, and need not involve any specific transaction of money among agents.

We assume that for each agent in C_m there is a sequence of times, at which the agent makes decisions. If t is a time when agent $a_i \in C_m$ is making a choice, then, we postulate that a_i switches to the WAN AP w_m from a hotspot h_k if and only if it was connected to h_k at time t^- and

$$U_i^w\left(N_m^w(t^-)+1\right) > U_i^h\left(N_k^h(t^-)\right) + c^w,$$

where t^- refers to the time immediately prior to t and c^w represents a cost of switching to the WAN AP. Similarly, the agent $a_i \in S_k^h$ switches to a hotspot h_k at time t if and only if it was connected to a WAN AP w_m at t^- and

$$U_i^h\left(N_k^h(t^-)+1\right) \geq U_i^w\left(N_m^w(t^-)\right) + c^h,$$

where c^h represents the cost of switching to a hotspot. Note that we break ties in favor of hotspots.

Assumption 2.4. *Agents' decision times within C_m are given by a simple point process Φ_m with realizations ϕ_m and assume that Φ_m obeys the following:*

- *Φ_m almost surely contains infinitely many points in $\mathbb{R}^+$, i.e. $\phi_m = \{s_k\}_{k=1}^{\infty}$, where $s_k \in \mathbb{R}^+$ for $k = 1, 2, \ldots$*
- *each point of ϕ_m is associated with a decision time for a unique agent within C_m*
- *a point $s_k \in \phi_m$ is a decision time of the agent $a_i \in C_m$ with some* **positive** *probability p_i, which possibly depends on realization ϕ_m up to time s_k and the history of agents choices up to time s_k.*

Assumption 2.4 postulates that only one agent within C_m can make decision at a time, each agent has unlimited opportunities for decision making, and any decision time with some positive probability is associated with a particular agent.

2.5 Analysis of decision dynamics

In this section we consider the dynamics of agents' decision making. In particular we investigate if the dynamics converge to a particular fixed point which we refer to as an "equilibrium". We define equilibrium similarly in the way Nash equilibria [48] are defined in game theory:

Definition 2.3. *Consider a service zone S_m^w of WAN AP w_m for a particular realization of agents, hotspots and WAN APs on the plane. We refer to a particular configuration of agents' choices within S_m^w as equilibrium, if, given this configuration, no agent will alter its choice.*

The rest of this chapter will be devoted to answering the following questions:

1. *Does the dynamics of greedy decision-making converge to a (fixed) equilibrium?*
2. *If the dynamics converge, what is the structure of the corresponding equilibria?*

This study is motivated by our ultimate goal, which aims to characterize the heterogenous wireless system in terms of (i) competitiveness of the two providers and (ii) end-users' performance. Guaranteeing that an equilibrium is eventually reached by the system will enable us to focus our attention on this equilibrium configuration, and associate the "competitiveness" and performance metrics with the structure of equilibrium.

Let us denote by M_m^w the total number of agents that fall within S_m^w for a particular realization. We will make the following assumption.

Assumption 2.5. *If $a_i, a_j \in \pi^a(S_k^h)$ where $k \in \mathcal{K}_m$ then $|U_i^w(N) - U_j^w(N)| < c^h + c^w$ for $1 \leq N \leq M_m^w$.*

Assumption 2.5 requires that the utility associated with connections to the WAN does not vary too much for agents located within the service zone of the same hotspot. For example if the performance of WAN connections simply degrades with distance, then this assumption requires the coverage radius of a single hotspot to be small enough.

Theorem 2.1. *Consider the service zone S_m^w for a particular fixed realization π^a, π^h and π^w. Then, under Assumptions 2.1–2.4, given any initial configuration of agents' choices, say, at time $t = 0$, the system converges a.s. to an equilibrium configuration as $t \to \infty$.*

The proof of Theorem 2.1 that we give below is built on analysis of the system "from first principles". Before plunging into the proof we elaborate on how our decision-making model compares to some existing ones that have been analyzed within the game-theoretic literature.

Potential games, congestion games and relationship to our model. It turns out that a simple variation (but not general form) of our model could be cast into the realm of the so-called "congestion games". A congestion game is any game where a collection of homogeneous agents have to choose from a finite set of alternatives, and where the payoff of a player depends on the number of players choosing each alternative. These were first introduced within the game-theoretic community[8] by Rosenthal [61] in 1973.

[8]However, as mentioned in [58], similar concepts have been developed within the transportation research community back in 1952, see, e.g. [59] and [60].

The simplest example of a congestion game is a routing game in a transportation network, in which there is a finite number N of cars that share M distinct roads. The car i has a finite number of strategies that it would choose from, and each of its strategies represents a subset $\mathcal{A}_i$ of roads that it would use (these roads would, for example form one of the alternative routes leading to the i's car destination). The cost of using road m, for $m = 1,\ldots,M$ is given by $c_m(N_m)$ for each car having route passing through this road; here N_m is the total number of cars sharing the road. (Note that $c(N_m)$ has a solely congestion dependent structure according to Definition 2.2.) All cars greedily minimize their cost, by selecting one of the choices available to them. Assuming only one car changes her strategy at a time leads to a time dynamics where congestion level at each road changes over time.

Convergence of such dynamics with time was established by Monderer and Shapley in [62] for a class of "potential games" that includes "congestion games" as a sub-class. The basic idea is to find a "potential" function which would depend on actions of all players and would be able to follow the changes in utilities of each individual player. For completeness, we will formally define potential games. The basic setup deals with games in "strategic form", which include:

- Finite set of players: $\mathcal{P} = \{1,2,\ldots,I\}$
- Finite set of (pure) strategies S_i available to player $i \in \mathcal{P}$. (We let $S \equiv S_1 \times S_2 \times \ldots \times S_I$ denote the Cartesian product of strategies' spaces. Also let $S_{-i} = S_1 \times S_2 \times \ldots \times S_{i-1} \times S_{i+1} \times \ldots \times S_I$, i.e. a set of strategies available to all players but the i-th player)
- A set of utility (payoff) functions $\{u_i\}_{i=1}^I$, where $\forall i \in \mathcal{P}$ we have $u_i : S \mapsto \mathbb{R}$. For example $u_i(s)$ denotes the utility function of player i when the strategies of all players are given by $s \in S$.

Definition 2.1. *A strategic form game G is a tuple*

$$\{\mathcal{P}, \{S_i\}_{i=1}^I, \{u_i\}_{i=1}^I\}$$

consisting of a set of players, pure strategy spaces and utility functions. A strategic form game is finite if strategy spaces S_i are finite for all $i \in \mathcal{P}$.

Definition 2.2. *A strategic form game is a potential game if there exists a function $\Phi : S \mapsto \mathbb{R}$, such that, for all $i \in \mathcal{P}$ and all $s_i \in S_i$, $s_{-i} \in S_{-i}$ we have:*

$$u_i(s_i, s_{-i}) - u_i(s_i', s_{-i}) = \Phi(s_i, s_{-i}) - \Phi(s_i', s_{-i}).$$

A path in S is a sequence $\Gamma = \gamma^0, \gamma^1, \ldots$, where $\gamma^n \equiv \{s_i^n : \ s_i^n \in S_i\}_{i\in\mathcal{P}}$, and, for every $n = 1,2,\ldots$, there exists a unique player $i(n)$, such that $\gamma^n = \{x, \gamma_{-i(n)}^{n-1}\}$ for some $x \neq s_{i(n)}^{n-1}$. γ_0 is

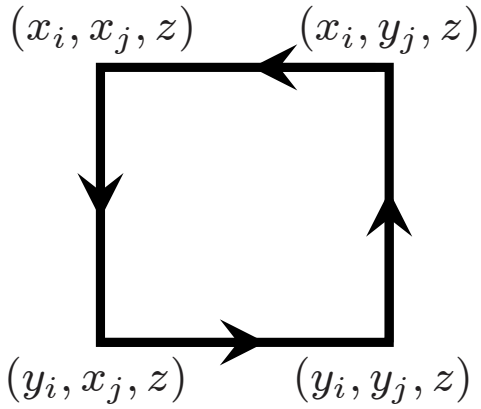


Figure 2.2: Closed loop $A \to B \to C \to D \to A$ in Proposition 2.3

called the initial point of Γ, and if Γ is finite, then its last element is called the terminal point of Γ. The path Γ in S is an improvement path if for all $n = 1, 2, \ldots$ we have $u_{i(n)}(\gamma_n) > u_{i(n)}(\gamma_{n-1})$, where $i(n)$ is the unique deviator at step n.

Proposition 2.1. [62] *Any improvement path in a finite potential game is finite.*

Proposition 2.2. [62] *If Φ is a potential function for a potential game G, then $s^* \in S$ is a (Nash) equilibrium of G if, for each i: $\Phi(s) \geq \Phi(x, s_{-i})$, i.e. $\Phi(s)$ is maximized over the unilateral deviations from s^*.*

Note that Propositions 2.1, 2.2 in fact imply that the Nash equilibria for each potential game are reached in a finite number of steps, under the assumption that only one player selects a better strategy per time step.

The following proposition gives a necessary and sufficient condition for a game to admit a potential function. Note that this condition resembles the corresponding requirement for a force field in physics to admit a representation via potential: the work of the force field over any closed loop must be zero.

Proposition 2.3. [62] *For any $x_i, y_i \in S_i$, and $x_j, y_j \in S_j$ and $z \in S_{-i-j}$ define $A = (x_i, x_j, z)$, $B = (y_i, x_j, z)$, $C(y_i, y_j, z)$ and $D = (x_i, y_j, z)$ (see Figure 2.2). Then a strategic form game G is a potential game if and only if:*

$$u_i(B) - u_i(A) + u_j(C) - u_j(B) + u_i(D) - u_i(C) + u_j(A) - u_j(D) = 0\,.$$

Corrollary 2.1. [62] *Any congestion game is a potential game.*

Corrollary 2.2. *For any congestion game, greedy decision-making strategies converge to Nash equilibria in a finite number of steps.*

Note that Corollary 2.2 establishes our Theorem 2.1 for the simple case where utilities of agents connected to both the WAN and WLANs are solely congestion-dependent. Indeed, our decision-making model represents a variant of a "congestion game". The highlight of this paragraph is that the general version of our model, however, does not belong to a class of congestion, or more generally does not belong to a class of potential games.

Proposition 2.4. *In our general model where utilities of agents connected to the WAN are congestion and agent dependent, the decision-making dynamics can not be represented by that of a congestion, or more generally, potential game.*

Proof. Note that strategy space S_i for each agent in our decision-making model consists of only two elements $\{h,w\}$ corresponding to agent selecting a hotspot or the WAN. We will now use Proposition 2.3 to test if the corresponding game admits representation via a potential function. Let us consider two agents $a_i, a_j \in S_m^w$ that fall into two different hotspots h_k and h_l (we assume $h_k, h_l \in S_m^w$). Let s_i, s_j, z denote a configuration in the strategy space where a_i made decision s_i and a_j made decision s_j, while all other agents within S_m^w keep with their decisions, which we will denote by z. Now consider a four distinct configurations in the space of agents' decisions: $A = \{h,h,z\}$, $B = \{w,h,z\}$, $C = \{w,w,z\}$, $D = \{h,w,z\}$, and the corresponding "closed loop" defined via transitions $A \to B \to C \to D \to A$. According to Proposition 2.3 the game would have a potential function iff:

$$\begin{aligned} &U_i^w(N_m^w+1) - U_i^h(N_k^h) + U_j^w(N_m^w+2) - U_j^h(N_l^h) \\ &\qquad + U_i^h(N_k^h) - U_i^w(N_m^w+2) + U_j^h(N_l^h) - U_j^w(N_m^w+1) \\ &= U_i^w(N_m^w+1) - U_j^w(N_m^w+1) + U_j^w(N_m^w+2) - U_i^w(N_m^w+2) = 0\,. \end{aligned} \tag{2.1}$$

Note that in (2.1) we assumed that N_m^w, N_k^h and N_l^h agents have been respectively connected to WAN AP w_m, hotspots h_k and h_l in configuration A. Clearly, $U_i^w(N) \neq U_j^w(N)$ in general, thus equation (2.1) is not satisfied and decision-making dynamics in our model does not represent a potential game. □

As discussed at length in [58], some convergence results exist for games that do not fall into class of potential games, but often particular scenarios have to be treated on a case by case basis. This motivates the approach that we took in proving Theorem 2.1, where we do not rely on any existing results from game theory, but instead construct a proof technique that works for our particular setup.

Proof of Theorem 2.1. To prove Theorem 2.1 we need the following technical lemma, that is a straightforward consequence of Assumption 2.5.

Lemma 2.1. *Consider $a_i, a_j \in \pi^a(S_k^h)$ where $h_k \in S_m^w$ and suppose Assumptions 2.3 and 2.5 hold. Furthermore let a_j be connected to h_k and suppose a_i switches from WAN AP w_m to h_k at time s. Then a_j can not switch from hotspot h_k to WAN AP at time $t > s$ if no other agent has switched in S_m^w during the time interval (s,t).*

Proof. We prove the lemma by contradiction. Assume that an agent a_j has switched from h_k to WAN AP w_m at time t, then the following condition must have been satisfied:

$$U_j^w\left(N_m^w(t^-)+1\right) > U_j^h\left(N_k^h(t^-)\right) + c^w\,. \tag{2.2}$$

Furthermore, suppose agent a_i switched from the WAN AP w_m to hotspot h_k at time $s < t$, and thus:

$$U_i^w\left(N_m^w(s^-)\right) + c^h \leq U^h\left(N_k^h(s^-)+1\right). \tag{2.3}$$

Since we have assumed that no other agent within S_m^w has switched in time interval (s,t), we have $N_m^w(s^-) = N_m^w(t^-)+1$ and $N_k^h(s^-) = N_k^h(t^-)-1$. Now combining (2.2) and (2.3), we have:

$$U_j^w\left(N_m^w(t^-)+1\right) - U_j^w\left(N_m^w(t^-)+1\right) > c^h + c^w\,,$$

that is in contradiction to Assumption 2.5. □

Below we give the essential details of the proof of Theorem 2.1. Note that under the assumptions of Theorem 2.1, the dynamics for the configuration of agents' decisions in S_m^w follow a continuous time Markov process with state $\mathcal{X}(t) := \{X(a_i,t)|\ a_i \in \pi^a(C_m)\}$, where $X(a_i,t) \in \{0,1\}$ – denotes the "connection state" of the agent a_i at time t and takes the value 0 if the agent is connected to a hotspot and 1 if it is connected to a WAN AP. (Note that we need only to consider the states of agents located within C_m.) We will classify transitions for this chain as "up", "down" and "stay", corresponding to agents switching from hotspots to the WAN AP, vice versa, or staying with their current choice. For simplicity we can "uniformize" the continuous-time process to focus on a discrete time Markov chain capturing times where decisions are made. We shall denote these decision times by $s = 1, 2, \ldots$. The transition probabilities for the discrete-time Markov chain are determined by two factors: the probability that a particular agent reconsiders her decision at that time, and whether the current configuration cause the agent to change providers.

By Assumption 2.1.4, each service zone contains an a.s. finite number of agents, thus there is an a.s. finite number of different configurations for agents' choices so the set of possible configurations is finite a.s.. It follows that some of the states must be revisited by the chain

infinitely often. To show the convergence of a system to an equilibrium, it is sufficient to construct a feasible path for the chain evolution which hits an equilibrium state with positive probability, *starting from any initial configuration*.

Below we present the steps of an algorithm to construct a path $\mathcal{P}$ consisting of a sequence of transitions for the state $\mathcal{X}(s)$, which, starting from any arbitrary configuration of agents' choices $\mathcal{X}(0)$, ends up in an equilibrium configuration after a finite number of steps. Let $A^u(s)$ denote the set of agents that, given the configuration at time s, could make "up" transitions and $A^d(s)$ the set of agents that can make "down" transitions. Let us also define a nondecreasing composite function,

$$J_i(N) \triangleq (U_i^w)^{-1} \circ (U_i^h(N) + c^w),$$

where $(U_i^w)^{-1}$ denotes a unique and decreasing, due to Assumption 2.3 inverse of U_i^w. We describe our algorithm in terms of pseudo-code shown in Table 2.2, where for convenience we denote $N^h(a_i,t) = N_k^h$, where k is such that $a_i \in S_k^h$. Note that our notational convention is that an agent making her decision at time slot $s \geq 1$ is basing this decision by observing the state of the system prior to that time, i.e. time $s-1$.

After initialization, the algorithm (see Table 2.2) alternates between the Up- and Down-transition phases. During the Up-transition phase only the "up"-switchings occur, where the agents performing these transitions are selected to be those which are the most "unsatisfied". This phase ends once the set of agents that are able to perform the "up"-transitions depletes. At that time the algorithm switches to the "down"-transition phase, where at most one agent performs a "down"-transition. We introduce an auxiliary integer sequence $\{Z(t)\}_{t=1}^{\infty}$ with values that depend on the state of the system prior to a transition at times $t = 1,2,\ldots$, and show that this sequence is nonincreasing. This allows us to argue that $Z(t)$ converges to a limit Z^* after an a.s. finite number of transitions. Then, we demonstrate that the equilibrium must be reached in a.s. finite time once $Z(t)$ has reached the level Z^*. The remaining steps of the proof are straightforward, but lengthy and are relegated to the Appendix of this chapter.

2.6 Structure of agents' choices equilibria

Note that, in general, the specific character of the agents' choices in equilibrium depends on the utility functions, distances from access points, and resource allocation mechanisms at the access points. For the service zone associated with a particular WAN AP an equilibrium might not even be unique. For solely congestion dependent utilities, however, in the next section we show that the set of all equilibria in each WAN service zone could be made quite "tight", under some conditions which we elaborate on below.

Initialization:
$s = 1$ and $\mathcal{X}(s) = \mathcal{X}(0)$
$Z(s) := 0$
go to Up-transition phase

Up-transition phase:
if $A^u(s) \neq \emptyset$
$\{\ j := \arg\max_{i:\ a_i \in \pi^a(A^u(s))} \left\lfloor J_i\left(N^h(a_i, s)\right)\right\rfloor$
$Z(s) := \left\lfloor J_j\left(N^h(a_j, s)\right)\right\rfloor$
let a_j make an "up" transition
update the state $\mathcal{X}(s)$
$s := s + 1\ \}$
otherwise: go to Down-transition phase

Down-transition phase:
if $A^d(s) \neq \emptyset$:
$\{$ choose any $a_j \in A^d(s)$
let a_j make a "down" transition
update the state $\mathcal{X}(s)$
$Z(s) := Z(s-1)$
$s := s + 1$
go to Up-transition phase $\}$
otherwise: done

Table 2.2: Pseudo-code for constructing the path $\mathcal{P}$ converging to equilibrium.

2.6.1 Solely congestion utilities for both WAN and WLAN connections

Let us consider the properties of the set of possible equilibria for a given service zone S_m^w associated with a particular WAN AP w_m. In this section we will assume that for each agent both utilities are solely congestion dependent, and the switching costs, $c^w = c^h = 0$. Moreover, to simplify exposition we will let the structure of utility for agents connected to different hotspots within S_m^w be the same, i.e., for all $N \in \mathbb{N}$, $U_i^h(N) = U_j^h(N)$ for agents a_i and a_j that belong to different hotspots within S_m^w. With such simplifications, we can drop the subscripts from utility functions that correspond to agents within S_m^w, in particular, with some abuse of notation we let $U^w(\cdot)$ to denote the utility of any agent within S_m^w connected to WAN AP, and $U^h(\cdot)$ denote the utility of any agent within S_m^w connected to a hotspot. Note that an example where this set of assumptions holds is a scenario where utilities of agents depend only on the share of bandwidth that they are able to get by connecting to a particular AP, i.e. all hotspots have the same available bandwidth and WAN and hotspots' APs split the available bandwidth equally among the agents connected to them.

Here we will give a characterization of the system state, i.e., configuration of agents' decisions, define a notion of uniqueness, and analyze under what conditions the system equilibrium is unique. We first introduce some additional notation[9]

- $M_m^w = \left|\pi^a(S_m^w)\right|$ – the number of agents located within the service zone of WAN AP w_m.
- $M_k^h = \left|\pi^a(S_k^h)\right|$ – the number of agents located within the service zone of hotspot h_k.
- $M_{C_m}^w = \left|\pi^a(C_m)\right|$ – the number of agents located within C_m, i.e. agents that can make choices.
- $M_{\bar{C}_m}^w = M_m^w - M_{C_m}^w$ – the number of agents located within $\bar{C}_m$, i.e. the agents that can *not* make choices.
- $H_m = \left|\pi^h(S_m^w)\right|$ – the number of hotspots located within the service zone of WAN AP w_m.
- $\mathcal{X}_m^{\pi^a,\pi^h,\pi^w} \triangleq \{X(a_i)|\ a_i \in \pi^a(C_m)\}$ – denotes the system configuration in service zone S_m^w associated with a fixed realization π^a, π^h and π^w. Here $X(a_i) \in \{0,1\}$ takes the value 0 if agent a_i is connected to a hotspot, and 1 if she is connected to a WAN AP.
- $\mathcal{T}_m = \mathcal{T}_m(\pi^a,\pi^h,\pi^w)$ – the a.s. finite set of possible system configurations (states) in S_m^w for a given realization π^a, π^h and π^w.
- $\mathcal{E}_m$ – the set of all system configurations $c \in \mathcal{T}_m$ that correspond to equilibria in S_m^w.
- $\mathcal{F}_m = \mathcal{F}_m(\pi^a,\pi^h,\pi^w)$ – subset of $\mathcal{E}_m$ which consists of only "fair" equilibria (see below).
- $N_m^w(c)$ – the number of agents that connect to WAN AP w_m in configuration $c \in \mathcal{T}_m$.
- $N_k^h(c)$ – the number of agents that connect to hotspot AP h_k in configuration $c \in \mathcal{T}_m$.
- $(U^w)^{-1}(\cdot)$ – unique and decreasing, by Assumption 2.3 inverse of $U^w(\cdot)$.
- $(U^h)^{-1}(\cdot)$ – unique and decreasing inverse of $U^h(\cdot)$.
- $G(\cdot) \triangleq (U^h)^{-1} \circ U^w(\cdot)$ – nondecreasing composition of $(U^h)^{-1}$ and $U^w(\cdot)$
- $J(\cdot) \triangleq (U^w)^{-1} \circ U^h(\cdot)$ – nondecreasing composition of $(U^w)^{-1}$ and $U^h(\cdot)$

[9]Note that we use letter M with different sub- and super- scripts to refer to the actual number of agents that fall within different sets, while we use the letter N to refer to the number of agents within different sets to refer to the agents *actually connected* to particular APs.

Characterization of a configuration. For any fixed realization π^a, π^h and π^w consider only WAN APs w_m that have at least one hotspot in their service areas, i.e. $\mathcal{K}_m \neq \emptyset$. For such m we characterize the system configuration $c \in \mathcal{T}_m$ for the service zone S_m^w by a vector $\mathbf{N}_m(c) \triangleq \{N_k^h(c) | \, k \in \mathcal{K}_m\}$. The vector $\mathbf{N}_m(c)$ determines how many agents are connected to each hotspot h_k for $k \in \mathcal{K}_m$ in configuration $c \in \mathcal{T}_m$.

Definition 2.3. *We say that two configurations for agents' choices characterized by* $\mathbf{N}_m(c)$ *and* $\mathbf{N}_m(c')$ *are equivalent, and write* $\mathbf{N}_m(c) \sim \mathbf{N}_m(c')$*, if the components of the vector* $\mathbf{N}_m(c)$ *are a permutation of those of* $\mathbf{N}_m(c')$.

Fair equilibria.

Definition 2.4. *We say that a configuration* $c \in \mathcal{T}_m$ *is "fair" if its characterization* $\mathbf{N}_m(c) = \{N_k^h(c) | k \in \mathcal{K}_m\}$ *satisfies, for some* $K \in \mathbb{Z}^+$*:*

$$\forall k \in \mathcal{K}_m : \begin{cases} K-1 \leq N_k^h(c) \leq K\,, & \text{if } M_k^h \geq K\,, \\ N_k^h(c) = M_k^h\,, & \text{otherwise}\,. \end{cases}$$

If c is also an equilibrium configuration we say that c is a "fair" equilibrium.

We shall interpret this definition via Figure 2.3. The hexagonal region is a schematic representation of the service zone S_m^w, while the positions of the cylinders represent the locations of hotspots. The height of each cylinder represents the overall number of agents that fall within the service zone of a particular hotspot.

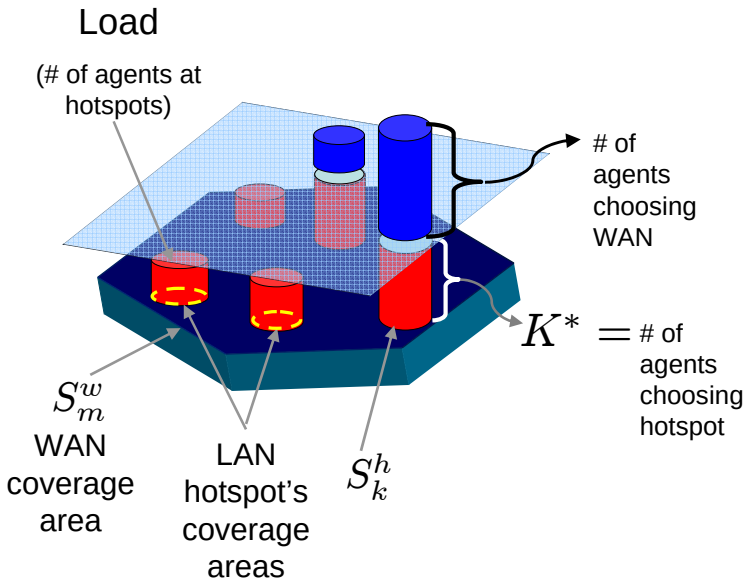


Figure 2.3: Structure of a "fair" configuration.

Assume that the slicing plane in Figure 2.3 is one unit thick and its upper surface is placed at integer-valued heights K above the surface of S_m^w. Any "fair" configuration has the following assignment of agents to APs:

- All agents in $S_m^w \setminus C_m$ connect to WAN AP w_m.
- A number of agents corresponding to the parts of cylinders that fall under the lower surface of the slicing plane connect to their respective hotspots.
- A number of agents corresponding to the parts of cylinders above the upper surface of the plane connect to the WAN AP w_m.
- Finally, a number of agents corresponding to the parts of cylinders within the slice connect to either their associated hotspots or WAN AP w_m.

In what follows, to avoid ambiguity we will always associate a fair configuration f with the "cut-off" plane at level[10] $K_m(f) = \max_{k \in \mathcal{K}_m} N_k^h(f)$. Note, that in fair configuration f the hotspots having more than $K_m(f)$ agents in their service zones yield the "overload" to the WAN AP w_m. As a result the number of agents connected to those hotspots is nearly the same, i.e. either $K_m(f)$ or $K_m(f) - 1$.

By the construction used to prove Proposition 2.5 we can always find a position of the slicing plane, $K = K_m^*$, and an assignment of agents corresponding to the parts of cylinders at the slice, so that the connection configuration in S_m^w is a fair equilibrium. This results in statement **(i)** of Proposition 2.5.

Proposition 2.5. *For any realization π^a, π^h and π^w we have that:*
(i) The set of all fair equilibria, $\mathcal{F}_m$, is not empty.
(ii) All fair equilibria have equivalent characterizations, i.e. for all $f, f' \in \mathcal{F}_m$, $\mathbf{N}_m(f) \sim \mathbf{N}_m(f')$.

For the proof of statement (ii) of Proposition 2.5, see Appendix 2.7.2.

Non-uniqueness of equilibrium.

Definition 2.5. *For a particular realization π^a, π^h and π^w we say that the equilibrium in S_m^w is unique if for any $e, e' \in \mathcal{E}_m$ we have $\mathbf{N}_m(e) \sim \mathbf{N}_m(e')$.*

Note that agents' decisions are discrete in nature, and unfortunately, this can lead to multiple equilibria in the system, even when we understand uniqueness in the weak sense of

[10]The ambiguity arises in the case when for a particular fair configuration $f \in \mathcal{F}_m$ we have $0 < N_k^h = K_m < \max_k M_k^h$, for all $k \in \mathcal{K}_m$ and for some $K_m \geq 0$. Then the upper surface of the "slicing plane", associated with this configuration can be drawn at either the levels K_m or $K_m + 1$.

Definition 2.5. Below we show this via a simple example. Observe that for all equilibrim configurations $e \in \mathcal{E}_m$ we must have that:

$$U^h\left(N_k^h(e)+1\right) < U^w\left(N_m^w(e)\right) \text{ and } U^h\left(N_k^h(e)\right) \geq U^w\left(N_m^w(e)+1\right) \tag{2.4}$$

for all $k \in \mathcal{K}_m$ such that the service zone S_k^h has an agent connected to the WAN AP w_m and an agent connected to h_k. Also we must have that:

$$U^h\left(N_l^h(e)\right) \geq U^w\left(N_m^w(e)+1\right),$$

for all $l \in \mathcal{K}_m$ such that the service zone S_l^h has *all* of its agents connected to h_l. Lastly,

$$U^h(1) < U^w\left(N_m^w(e)\right), \tag{2.5}$$

must be satisfied for all $p \in \mathcal{K}_m$ such that all agents within S_p^h are connected to w_m in equilibrium. It follows from (2.4) that:

$$G\left(N_m^w(e)\right) - 1 < N_k^h(e) \leq G\left(N_m^w(e)+1\right), \tag{2.6}$$

for hotspots $h_k \in \pi^h(S_m^w)$ with at least one agent connected to the WAN AP w_k. Note that, depending on the utility functions there can be more than one integer solution to the inequalities (2.6). Consider, for example:

$$U^h(N) = N^{-\beta}, \quad U^w(N) = N^{-\alpha}, \tag{2.7}$$

where $\alpha > \beta > 0$. In this case $G(N) = N^{\alpha/\beta}$, and the gap between the right and left hand side in (2.6) increases with $N_m^w(e)$. In other words, when the number of agents, not covered by hotspots is large enough, there can be many integer solutions to the inequalities (2.6). Hence, "unfair" equilibria can be constructed easily from the fair one. For example we could switch some number L of agents from WAN AP w_m to a particular hotspot h_k and the same number L of agents from some other hotspot h_l within the same WAN AP w_m. Note that this procedure would not change the number of agents connected to the WAN AP. If L is selected so that the $N_k^h(e) - L$ and $N_l^h(e) + L$ are still within the bounds (2.6), this procedure would result in a feasible equilibrium which is not equivalent to the fair one.

Conditions guaranteeing uniqueness and fairness. One might ask under what conditions the equilibrium in S_m^w is unique. The following result assumes that the utilities have a particular property, and that the cells of the WAN provider are large enough to guarantee that a sufficiently large number of agents connects to the WAN AP in equilibrium.

Proposition 2.6. *Suppose that there exists $\overline{N}$ such that for all $N \geq \overline{N}$*

$$G(N+1) - G(N) < 1\,, \tag{2.8}$$

and assume that the number of agents that can not make choices in service zone S_m^w satisfies:

$$M_{\bar{C}_m}^w \geq \overline{N}\,. \tag{2.9}$$

Then, the equilibrium in S_m^w is unique and fair.

We prove this proposition in Appendix 2.7.3. In general, if the property (2.8) holds then it must be the case that the utility function associated with connections to hotspots decrements faster in the number of connected agents than the utility associated with connections to the WAN AP[11]. One such example is given by (2.7) with $\beta > \alpha > 0$.

Max-min fairness property of unique equilibria. Let us define $U_m^{min}(c)$ to be the minimum over the utilities of agents within S_m^w that choose according to configuration $c \in \mathcal{T}_m$. We refer to $U_m^{min}(c)$ as the utility of the bottleneck agent for the configuration c. The following proposition, which we prove in Appendix 2.7.4, implies that when equilibrium configuration of agents' choices is unique, it would maximize the utility of the bottleneck agents over all possible configurations of agents' choices. This property will become crucial when we consider efficient heterogeneous network design problems in Chapter 4.

Proposition 2.7. *If the equilibrium in S_m^w is unique, then $U_m^{min}(c) \leq U_m^{min}(f)$, for all $f \in \mathcal{F}_m$ and $c \in \mathcal{T}_m$.*

2.6.2 Structure of equilibrium in a more general setting.

We now return to our general model where the utility functions of agents connected to the WAN are of congestion and agent dependent types. It is easy to identify the conditions for equilibrium corresponding to the service zone S_m^w. Let N_k^h be the equilibrium number of agents within S_k^h connected to h_k, and denote by $\mathcal{W}_m(k)$ and $\mathcal{H}_k$ the indices of agents located within S_k^h and connected to WAN AP w_m and hotspot h_k respectively. Then using equilibrium conditions for all agents within S_k^h it is easy to establish:

$$\max_{i \in \mathcal{W}_m(k)} \left[(U_i^h)^{-1} \circ (U_i^w(N_m^w) + c^h) \right] - 1 < N_k^h \leq \min_{j \in \mathcal{H}(k)} \left[(U_j^h)^{-1} \circ (U_j^w(N_m^w + 1) - c^w) \right]\,, \tag{2.10}$$

[11] Since, as we alluded above, the WAN service might be degrading slower with the number of connections than that of hotspots, the assumption that (2.8) holds may be reasonable.

where we assumed the convention that max over empty set is equal to -1 (lower bound is not binding when no agents within S_k^h are connected to the WAN) and the convention that min over empty set is equal to ∞ (upper bound is not binding when no agents are connected to hotspot h_k).

From the complex expression (2.10) it can be seen that equilibria for our general model are much harder to identify and their structure is more complex. A simple case is that where utilities of agents connected to the WAN and located within the range of the same hotspot are close in value – in this case the structure of the equilibria resembles Figure 2.3, where the slicing surface has a spatially dependent structure with both its level and thickness varying from one hotspot to another.

To provide a sensible illustration we will study equilibria for a particular scenario which takes Example 2.1 as a basis. We consider a service zone of a WAN AP w_m populated with M_m^w agents that are trying to maximize their average over time throughput, thus the utility of an agent $a_i \in S_m^w \cap S_k^h$, when it is connected to the WAN is

$$U_i^w(N) = \frac{B_i^w}{N} \tag{2.11}$$

and when it is connected to the hotspot is:

$$U_i^h(N) = \frac{B_k^h}{N}. \tag{2.12}$$

To make utility $U_i^w(\cdot)$ congestion and agent dependent, but same for all agents residing within the same hotspot, we assume that B_i^w depends on location of agent's a_i hotspot, h_k. For example, we can have (for downlink transmissions):

$$B_i^w = \frac{C}{2}\log\left(1 + \frac{P_m^w L_m^w(h_k)}{\Gamma \mathcal{N}_0}\right) \tag{2.13}$$

where C is the size of a frequency band used for WAN transmissions, P_m^w is the (fixed) power level used at WAN AP, $L_m^w(h_k)$ is the path loss from w_m to h_k (depending on shadowing, large scale attenuation, etc.), $\mathcal{N}_0$ is the ambient noise power level at the agent's receiver and $\Gamma \geq 1$ is the "gap" which depends on modulation scheme. To make utility $U_i^h(\cdot)$ solely congestion dependent we require B_k^h to be independent of i – this would model quite well a typical [8] scenario where wireless access bandwidth at, e.g., Wi-Fi hotspots exceeds by far the available to hotspots backhaul bandwidth.

Using the equilibrium conditions (2.10) and expressions (2.11,2.12) for utilities we ob-

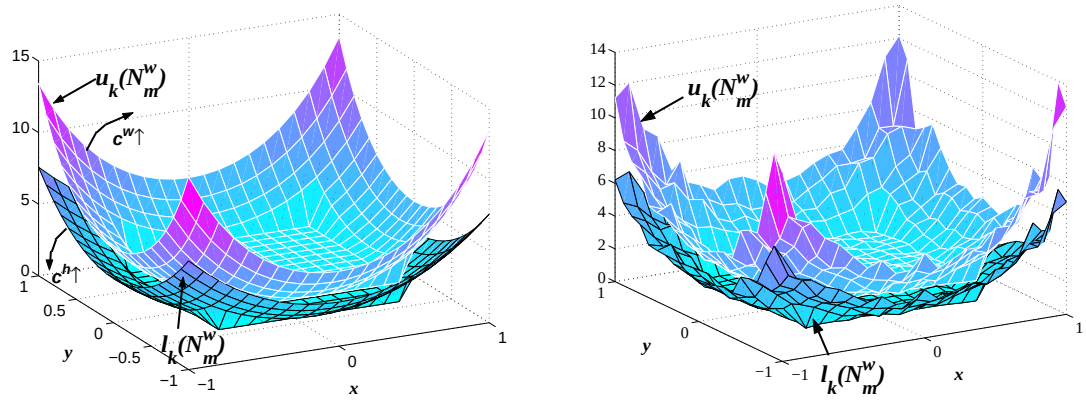


Figure 2.4: Spatial structure of the "slicing surface" when $L_m^w(h_k)$ has spatially dependent mean given by (2.15) with $\alpha = 3$. (a) log-normal component in attenuation is small (b): Log-normal component in attenuation is significant.

tain:

$$l_k(N_m^w) := \frac{B_k^h}{\frac{B_i^w}{N_m^w} + c^h} - 1 < N_k^h \leq \frac{B_k^h}{\left[\frac{B_i^w}{N_m^w+1} - c^w\right]^+} := u_k(N_m^w)\,. \tag{2.14}$$

Expression (2.14) provides a lower and upper bounds for the number of agents connected to hotspot h_k in equilibrium given that N_m^w agents are connected to the WAN AP. Note that the bounds might become loose, when, e.g. $l_k < 0$ or $u_k = \infty$. These bounds correspond to lower and upper envelops of the "slicing surface" which replaces the plane in Figure 2.3. Note that l_k and u_k both depend on B_i^w, and hence, on the strength of the signal from the WAN AP at location h_k, thus the shape of both envelopes is spatially dependent, as well as the gap between them.

In Figure 2.4 we show the spatial structure of $l_k(N_m^w)$ and $u_k(N_m^w)$, for a fixed value of $N_m^w = 10$. The values of x and y are Cartesian coordinates of the hotspot h_k, and we assume that the location of the WAN AP w_m is at the center of this coordinate system. The path loss $L_m^w(h_k)$ for each h_k is lognormally distributed with spatially dependent average given by simple power-law decay function

$$\bar{L}_m^w(h_k) = \min\left(A, \frac{1}{|h_k - w_m|^\alpha}\right), \tag{2.15}$$

and the backhaul bandwidth B_k^h is equal for all $k \in \mathcal{K}_m$. In Figure 2.4 (a) we plot the case where log-normal deviations from the average $\bar{L}_m^w(h_k)$ are negligible and, for comparison, in Figure 2.4 (b) the case where these fluctuations are more visible.

In Figure 2.4 we also indicate the tendency of the gap between the upper and lower envelopes to increase with the switching costs c^w and c^h. The gap also increases with distance of a hotspot from the WAN AP. Indeed, according to (2.14) the difference between $l_k(N_m^w)$ and

$u_k(N_m^w)$ is:

$$u_k(N_m^w) - l_k(N_m^w) \geq 1 + \frac{B_k^h}{B_i^w},$$

and thus this difference will most likely increase with distance from WAN AP, as long as B_k^h is close for all $h_k \in S_m^w$ and B_i^w degrades with distance.

One obvious consequence of the increasing distance between lower and upper envelopes of the slicing surface, is that the equilibria can realize in a number of different ways. This means unique equilibria will in general be an unlikely property for such a system indicating that most of the results of Section 2.6 will no longer hold in this more general setting.

2.7 Appendix to Chapter 2

2.7.1 Remainder of the proof of Theorem 2.1

Note that if the algorithm does not enter an Up-transition phase then there can only be "down" transitions in the system. Since the number of agents that are connected to each WAN AP is finite, the system will inevitably converge to an equilibrium which has no agents connected to the WAN AP w_m. Instead, assume that the system enters the Up-transition phase at time t_0. We will show that the sequence $Z(s)$, $s = t_0, t_0+1, \ldots$, defined in Table 2.2, is a non-increasing sequence.

We start by relating the function $J_j(\cdot)$ to agent a_j's eligibility for an "up" transition at time t. We must have:

$$U_j^w\Big(N_m^w(t-1)+1\Big) > U_j^h\Big(N^h(a_j,t-1)\Big) + c^w,$$

for an agent to be eligible to switch "up" at time $t \geq 1$. This is equivalent to:

$$N_m^w(t-1) < J_j\Big(N^h(a_j,t-1)\Big) - 1, \tag{2.16}$$

which in turn can be strengthened to:

$$N_m^w(t-1) \leq \Big\lfloor J_j\Big(N^h(a_j,t-1)\Big)\Big\rfloor - 1 \tag{2.17}$$

with a strict inequality in (2.17) if $J_j\Big(N^h(a_j,t-1)\Big) \in \mathbb{N}$.

Now consider any Up-transition phase. Note that $\Big\lfloor J_i\Big(N^h(a_i,s)\Big)\Big\rfloor$ can only decrease for each agent $a_i \in S_m^w$. Indeed, for each i, the function $J_i(\cdot)$ is nondecreasing and $N^h(a_i,s)$ could only be reduced during an Up-transition phase. Now, since the number of agents connected to the WAN AP w_m could only increase and by Assumption 2.3, $U_i^w(\cdot)$ is a decreasing function,

the value $U_i^w(N_m^w(s))$ can only decrease during an Up-transition phase. Clearly, by the general eligibility requirement (2.17), we have that the set of agents eligible for "up" transitions can only diminish within the Up-transition phase. Hence $A^u(s_1) \subset A^u(s_2)$, when $s_1 < s_2$ are both restricted to the period of the same Up-transition phase. Thus, for such s_1 and s_2:

$$\begin{aligned} Z(s_1) &= \max_{i:a_i \in A^u(s_1)} \left\lfloor J_i(N^h(a_i, s_1)) \right\rfloor \\ &\geq \max_{i:a_i \in A^u(s_2)} \left\lfloor J_i(N^h(a_i, s_2)) \right\rfloor = Z(s_2)\,, \end{aligned}$$

and hence $Z(s)$ is a nonincreasing sequence whenever s is within a single Up-transition phase.

We now show that $Z(s)$ is in fact nonincreasing for all $s \geq t_0$. Suppose that an Up-transition Phase finished at time $\tau+1$, and a_j was the agent that switched "up" at time τ, hence $Z(\tau) = \left\lfloor J_j\Big(N^h(a_j, \tau-1)\Big) \right\rfloor$. We will consider two scenarios. In the first scenario there is only one "down" transition at time $\tau+1$ and $A^u(\tau+2)$ becomes nonempty. We will show that in this scenario $Z(\tau+2) \leq Z(\tau)$. In the second scenario there is a sequence of $n > 1$ "down" transitions, before the set $A^u(\tau+n+1)$ becomes nonempty for the first time. In this scenario we will show once again that $Z(\tau+n+1) \leq Z(\tau)$.

Scenario 1: $A^u(\tau+2) \neq \emptyset$**.** Observe that once an agent a_i has performed a "down" transition at time $\tau+1$, we have:

$$J_k\Big(N^h(a_k, \tau+1)\Big) = J_k\Big(N^h(a_k, \tau)\Big)\,,$$

for all agents $a_k \in S_m^w$ that do not fall within the service zone of the same hotspot as a_i. For such agents we also have that:

$$\left\lfloor J_k\Big(N^h(a_k, \tau)\Big) \right\rfloor \leq \left\lfloor J_j\Big(N^h(a_j, \tau)\Big) \right\rfloor = Z(\tau)\,,$$

since a_j was chosen to make an "up" transition at time τ. Hence we have that for each agent that does not fall within service zone of the same hotspot as a_i:

$$\left\lfloor J_k\Big(N^h(a_k, \tau+1)\Big) \right\rfloor \leq \left\lfloor J_j\Big(N^h(a_j, \tau)\Big) \right\rfloor = Z(\tau)\,. \tag{2.18}$$

Now, by Lemma 2.1, no agent a_k that falls in the service zone of the same hotspot as a_i could switch "up" immediately after a_i has switched "down", and thus $a_k \notin A^u(\tau+2)$. But then, in view of (2.18) and the definition for $Z(t)$, we conclude that:

$$Z(\tau+2) \leq Z(\tau).$$

Scenario 2: $A^u(\tau+l)=\emptyset$ **for** $l=1,\dots n$ **and** $A^u(\tau+n+1)\neq\emptyset$**.** We will show that

$$Z(\tau+n+1)\leq Z(\tau)\,, \tag{2.19}$$

by contradiction. Assume that the inequality (2.19) is not satisfied. Then, we must have that:

$$\left\lfloor J_k\Big(N^h(a_k,\tau+n-1)\Big)\right\rfloor > \left\lfloor J_j\Big(N^h(a_j,\tau)\Big)\right\rfloor, \tag{2.20}$$

for some agent a_k within C_m. Indeed, consider an agent a_i that switches "down" at time $\tau+n$. Since a_i's switching down does not affect the number of agents connected to hotspots that do not contain a_i in their service zone, we have:

$$J_r\Big(N^h(a_r,\tau+n-1)\Big)=J_r\Big(N^h(a_r,\tau+n)\Big)\,,$$

for all agents $a_r\in S_m^w$ that do not fall within the service zone of the same hotspot as a_i. Moreover, by Lemma 2.1, no agent a_p that belongs to the service zone of the same hotspot as a_j can be eligible for an "up" transition at time $\tau+n+1$, i.e. $a_p\notin A^u(\tau+n+1)$. Hence, if

$$\begin{aligned}\max_{l:a_l\in A^u(\tau+n+1)}&\left\lfloor J_l\Big(N^h(a_l,\tau+n)\Big)\right\rfloor\\ &=Z(\tau+n+1)>Z(\tau)\,,\end{aligned}$$

then

$$\max_{l:a_l\in S_m^w}\left\lfloor J_l\Big(N^h(a_l,\tau+n-1)\Big)\right\rfloor > Z(\tau)\,,$$

which translates into (2.20).

Next we show that the agent a_k, where k satisfies (2.20), was eligible to switch "up" at time $\tau+n$. Consider again an agent a_j that switched "up" at time τ. To be eligible for making an "up" switch at time τ, according to (2.17) we must have:

$$N_m^w(\tau-1)\leq\left\lfloor J_j\Big(N^h(a_j,\tau-1)\Big)\right\rfloor-1 \tag{2.21}$$

with a strict inequality in (2.21) if $J_j\Big(N^h(a_j,\tau-1)\Big)\in\mathbb{N}$. Now consider the agent a_k, and note that

$$N_m^w(\tau+n-1)\leq N_m^w(\tau-1)\,, \tag{2.22}$$

since one agent has switched "up" at time τ and at least one agent has switched "down" at time interval $(\tau,\tau+n-1]$. Considering the agent a_k at time $\tau+n-1$ in view of (2.20), (2.21)

and (2.22) we obtain

$$N_m^w(\tau+n-1) < \left\lfloor J_k\Big(N^h(a_k,\tau+n-1)\Big)\right\rfloor - 1\,.$$

This leads to:

$$U_k^w\left(N_m^w(a_k,\tau+n-1)+1\right) > U^h\Big(N^h(a_k,\tau+n-1)\Big)\,,$$

and hence the agent a_k was eligible for an "up" transition at time $\tau+n$. We thus have a contradiction with the assumption that no agents were eligible for "up" transitions in the interval $(\tau,\tau+n]$. This proves that the inequality (2.19) holds.

To summarize we have proved that:

1. $Z(s)$ is nondecreasing if s is restricted to the period of a single Up-transition phase
2. If $\tau+1$ is the time when an Up-transition phase has finished and $\tau+n+1$, for $n \geq 1$ is the time when the next Up-transition phase has started, then: $Z(\tau) \geq Z(\tau+n+1)$.

Therefore $Z(s)$ is a nonincreasing sequence and since $Z(s)$ is integer valued it must have an integer-valued limit Z^* which is reached by the sequence in a.s. finite time t_1. We are left to show that the algorithm needs a.s. finite number of steps before an equilibrium is in fact reached.

For $t \geq t_1$, we have $Z(t) = Z^*$ and there are three scenarios for system evolution. The first scenario corresponds to the case where only "down" transitions take place in the system, and in the second – only "up" transitions are possible. In both of these scenarios the system reaches an equilibrium once all agents in C_m have switched to either the WAN or their respective hotspots. The third scenario is when the system undergoes both "up" and "down" transitions that are intermingled and we consider this scenario below.

From (2.17) we obtain that the agent is eligible for an "up" transition at time $t > t_1$ if and only if:

$$J_j(N^h(a_j,t-1)) = Z^* + \eta\,, \tag{2.23}$$

where $\eta \in [0,1)$, and

$$N_m^w(t-1) \leq Z^* - 1\,, \tag{2.24}$$

with strict inequality if $J_j(N^h(a_j,t-1)) \in \mathbb{N}$. Now, assume that an Up-transition phase has ended at time $\tilde{\tau}+1 > t_1$. Then this phase could finish either of the conditions (2.23) or (2.24) or both were violated at time $t = \tilde{\tau}+1$ for all agents $a_j \in C_m$.

Suppose that at time $t = \tau+1$ inequality (2.23) is violated for all $a_j \in C_m$, but inequality (2.24) is not. It is sufficient to show that no agent can become eligible for an "up" transition ever again at times $t \geq \tilde{\tau}+1$, since then the algorithm exits once all agents in C_m have connected

to their respective hotspots. To show this, we note first that if for any agent $a_j \in C_m$

$$J_j(N^h(a_j,\tau)) \geq Z^* + 1 \tag{2.25}$$

then, from (2.24)

$$N_m^w(\tau) < J_j(N^h(a_j,\tau)) - 1\,,$$

which indicates, via (2.16) that a_j is eligible to switch "up" at time $\tilde{\tau}$. This is in contradiction to what we assumed in the beginning of the paragraph and hence we can only have:

$$J_j(N^h(a_j,\tau)) \leq Z^* - 1\,, \tag{2.26}$$

for any agent $a_j \in C_m$. Now observe that, by (2.23) an agent $a_j \in C_m$ may become eligible for an "up" transition at time $t > \tilde{\tau}+1$ only if $J_j(N^h(a_j,t))$ has increased to or above level Z^*. Since $J_j(\cdot)$ is nondecreasing, there must be a "down" transition that would occur within the hotspot that contains a_j in its service zone. But if such "down" transition happens at time $\tilde{\tau}+1$, no agent within the service zone of a hotspot containing a_j can become eligible for an "up" transition at time $\tilde{\tau}+2$ as we proved in Lemma 2.1. Clearly we have that (2.24) is still satisfied for $t = \tilde{\tau}+1$, but then we must have

$$J_j(N^h(a_j,\tilde{\tau}+1)) < Z^*\,, \tag{2.27}$$

for any $a_j \in C_m$ since no agent is still available for an "up" transition. By induction we can prove that no agent is eligible for an "up" transition at any time $t \geq \tilde{\tau}+2$.

Now suppose that (2.24) is violated at time $t = \tilde{\tau}+1$. Note that the condition (2.24) was met at time $t = \tilde{\tau}$, since an "up" transition occurred at time $\tilde{\tau}$. Due to this "up" transition, we also have $N_m^w(\tilde{\tau}) = N_m^w(\tilde{\tau}-1)+1$ which yields $N_m^w(\tilde{\tau}) = Z^*$. Now let $\mathcal{L}(\tilde{\tau}+1)$ denote the set of agents for which:

$$\left\lfloor J_k(N^h(a_k,\tilde{\tau})) \right\rfloor = Z^*\,. \tag{2.28}$$

If no "down" transition occurs at time $\tilde{\tau}+1$ then the algorithm has exited and thus an equilibrium has been reached. Otherwise, assume that the "down" transition at time $\tilde{\tau}+1$ was in the service zone S_l^h of a hotspot $h_l \in S_m^w$. Note, that by Lemma 2.1, no agent that falls within S_l^h can become eligible for an "up" transition at time $\tilde{\tau}+2$, thus, since $N_m^w(\tilde{\tau}+1) = Z^*-1$, the condition (2.23) must be violated for those agents at time $\tilde{\tau}+2$. Hence these agents could not be within the set $\mathcal{L}(\tilde{\tau}+1)$ since otherwise they would be eligible for an "up" transition at time $\tau+2$. Furthermore, transition to S_l^h does not change the number of agents connected to hotspots $h_n \neq h_l$, and thus we have:

$$J_i(N^h(a_i,\tilde{\tau}+1)) = J_i(N^h(a_i,\tilde{\tau}))$$

for $a_i \notin S_l^h$. We thus conclude that $\mathcal{L}(\tilde{\tau}+1) = \mathcal{L}(\tilde{\tau}+2)$.

Now if $\mathcal{L}(\tilde{\tau}+1)=\emptyset$ then, similarly to as we argued above, no agent can ever become eligible for an "up" transition and the algorithm exits in finite time. Otherwise, if $\mathcal{L}(\tilde{\tau}+1)\neq\emptyset$ then at least one agent $a_j\in\mathcal{L}(\tilde{\tau}+2)$ is eligible for an "up" transition at time $\tilde{\tau}+2$ by sufficient conditions (2.23)-(2.24), since $N_m^w(\tilde{\tau}+1)=Z^*-1$ and $\lfloor J_j(N^h(a_j,\tilde{\tau}+1))\rfloor=Z^*$. Thus a_j can perform an "up" transition, which can only diminish the set $\mathcal{L}$ at a subsequent time $\tilde{\tau}+3$. By induction we thus can show that the set $\mathcal{L}(t)$ necessarily depletes in finite time, whence the algorithm exits.

In summary we have shown that from any starting configuration there exists a path, that with positive probability reaches an equilibrium state. Since the state space is finite, there must be a state which is visited infinitely often. Whence the Markov chain will necessarily eventually hit an equilibrium state.

2.7.2 Proof of Proposition 2.5

Proof. Consider any fair equilibrium configuration $f\in\mathcal{F}_m$ and let $K_m(f)=\max_{k\in\mathcal{K}}N_k^h(f)$ give the level of the corresponding slicing plane (see Figure 2.3). We will first show that for any two fair equilibria f and f' we have that $K_m(f)=K_m(f')$.

We show this by contradiction, suppose, in fact that there exist $f,f'\in\mathcal{F}_m$ such that $K_m(f)\neq K_m(f')$. Without loss of generality assume that $K_m(f)>K_m(f')$. Note that in this case for some $l\in\mathcal{K}_m$ we have $N_l^h(f)=K_m(f)\geq 1$. Considering the hotspot h_l, we get

$$U^h\Big(K_m(f)\Big)\geq U^w(N_m^w(f)+1) \tag{2.29}$$

since otherwise an agent connected to this hotspot would choose to switch to WAN AP w_m which would contradict the fact that f is an equilibrium. Now, for equilibrium f' all hotspots have fewer than or equal to $K_m(f')\leq K_m(f)-1$ agents, so in particular $N_l^h(f')\leq K_m(f)-1$. It follows by adding 1 to both sides and the fact that $U^h()$ is monotonically decreasing that:

$$U^h(N_l^h(f')+1)\geq U^h\Big(K_m(f)\Big). \tag{2.30}$$

At the same time, since $K_m(f')<K_m(f)$ it follows that $N_m^w(f')\geq N_m^w(f)+1$. Using the fact that $U^w()$ is monotonically decreasing we have that

$$U^w(N_m^w(f)+1)\geq U^w(N_m^w(f')). \tag{2.31}$$

Now putting (2.29),(2.30) and (2.31) together we have that

$$U^h(N_l^h(f')+1)\geq U^w(N_m^w(f'))$$

which implies that under f' an agent on WAN AP w_m would choose to switch to hotspot h_l. This contradicts the fact that f' is an equilibrium. Thus we conclude that for any $f \in \mathcal{F}_m$ we have $K_m(f) = K_m^*$ for some integer K_m^*.

In order to show that all fair equilibria are equivalent, we first argue that for two fair equilibria $f' \neq f$ we must have $N_m^w(f) = N_m^w(f')$. Without loss of generality suppose $N_m^w(f') \geq N_m^w(f) + 1$. Then, for at least one hotspot, say h_l, $N_l^h(f') \leq N_l^h(f) - 1$ which also implies that $N_l^h(f) \geq 1$. For f to be an equilibrium we must have that:

$$U^h\Big(N_l^h(f)\Big) \geq U^w(N_m^w(f) + 1) \geq U^w(N_m^w(f'))\,, \tag{2.32}$$

which follows from the fact that no agent in hotspot h_l wishes to switch to the WAN AP and our assumption. Considering the hotspot h_l under the equilibrium configuration f' we obtain:

$$U^w(N_m^w(f')) > U^h\Big(N_l^h(f') + 1\Big) > U^h(N_l^h(f))\,, \tag{2.33}$$

which is the consequence of the fact that an agent in h_l connected to the WAN AP w_m has no desire to switch to the hotspot h_l. Clearly, by monotonicity of utilities we have that (2.32) is in contradiction to (2.33).

Thus we know that if $f, f' \in \mathcal{F}_m$, then we have $N_m^w(f) = N_m^w(f')$ and $K_m(f) = K_m(f') = K_m^*$, for some integer K_m^*. Next we show that all fair equilibria must have equivalent characterizations. Let R denote the number of hotspots in S_m^w that have at least $K_m^* - 1$ agents in their service zones. The equilibrium number of agents connected to such hotspots is between $K_m^* - 1$ and K_m^*. Now assume that $r < R$ of the R hotspots have $K_m^* - 1$ agents and the remaining $R - r$ hotspots have K_m^* agents, connected to their APs under the equilibrium configuration f. Similarly, we assume that $r' < R$ hotspots have $K_m^* - 1$ agents in the equilibrium configuration f'. Equating the total number of agents in the service zone S_m^w in equilibria f and f', we have that:

$$\begin{aligned}(K-1)r + K(R-r) + \sum_{k \in \mathcal{K}_m,\ M_k^h < K_m^* - 1} M_k^h + N_m^w(f) \\ = (K-1)r' + K(R-r') + \sum_{k \in \mathcal{K}_m,\ M_k^h < K_m^* - 1} M_k^h + N_m^w(f')\,.\end{aligned}$$

Since $N_m^w(f) = N_m^w(f')$ this leads to $r = r'$, showing that $\mathbf{N}_m(f) \sim \mathbf{N}_m(f')$. □

2.7.3 Proof of Proposition 2.6

Proof. By part (i) of Proposition 2.5 there exists a fair equilibrium in S_m^w. Let $f \in \mathcal{F}_m$ be one such equilibrium and let $K_m(f) = \max_{k \in \mathcal{K}_m} N_k^h(f)$. We will consider three cases based on the value of $K_m(f)$ and show that under the assumptions of the proposition, any other equilibrium,

$e \in \mathcal{E}_m$ has the same characterization.

Case 1: $K_m(f) = 0$ In this case there is no agent in S_m^w which connects to a hotspot. If there are no agents within any of the hotspots' service zones, then it is nothing to prove, since no agents make any choices. Otherwise, considering the equilibrium conditions for agents that fall within some hotspot we have:

$$U^w(M_m^w) > U^h(1)\,. \tag{2.34}$$

It follows that no other equilibrium configuration can exist. Indeed, if $e \neq f$ is some other equilibrium configuration, we must have $N_l^h(e) \neq N_l^h(f)$, and thus $N_l^h(e) \geq 1$ yielding $N_m^w(e) \leq M_m^w - 1$. By Assumption 2.3 on utilities, we obtain:

$$U^w\left(M_m^w\right) \leq U^w\left(N_m^w(e)+1\right) \text{ and } U^h\left(N_l^h(e)\right) \leq U^h(1)\,. \tag{2.35}$$

Since e is an equilibrium, we should have:

$$U^w\left(N_m^w(e)+1\right) \leq U^h\left(N_l^h(e)\right), \tag{2.36}$$

since no agent in S_l^h wishes to switch to WAN AP w_m. Combining inequalities (2.35) and (2.36) we obtain:

$$U^w\left(M_m^w\right) \leq U^h(1)\,,$$

which contradicts inequality (2.34).

Case 2: $0 < K_m(f) = \max_{k\in\mathcal{K}_m} M_k^h$ In this case we have that there are no agents in C_m connected to the WAN AP w_m in configuration e and thus we have $\mathbf{N}_m(f) = \{M_k^h | k \in \mathcal{K}_m\}$. This can only be feasible if:

$$U^w\left(M_{\bar{C}_m}^w\right) \leq U^h(M_k^h)\,,$$

for $k \in \mathcal{K}_m$. Using this inequality instead of (2.34) and following the steps similar to the Case 1 one can prove that no equilibrium e exists, such that $N_k^h(e) < M_k^h$ for some $k \in \mathcal{K}_m$.

Case 3: $0 < K_m(f) < \max_{k\in\mathcal{K}_m} M_k^h$ Consider any other equilibrium $e \neq f$ and note that $N_m^w(e) \geq M_{\bar{C}_m}^w$. Hence the inequalities (2.6) admit at most two integer solutions. It follows that, for some $K \geq 1$ we have that:

$$K - 1 \leq N_k^h(e) \leq K\,,$$

for $k \in \mathcal{K}_m$ such that $M_k^h \geq K$ and

$$N_k^h(f) = M_k^h\,,$$

otherwise. Hence e must be a fair equilibrium, characterized by the slicing plane at level $K_m(e) = K$. Since by part **(ii)** of Proposition 2.5, all fair equilibria are equivalent, we have, that $\mathbf{N}(e) \sim \mathbf{N}(f)$. □

2.7.4 Proof of Proposition 2.7

For any configuration $c \in \mathcal{T}_m$ we will refer to agents that have utility equal $U_m^{min}(c)$ as the "bottleneck" agents. Let $c \in \mathcal{T}_m$ be a configuration that maximizes utility of a bottleneck agent and $c \notin \mathcal{F}_m$. We will show that $U_m^{min}(c) \leq U_m^{min}(f)$, for all $f \in \mathcal{K}_m$. Since, by assumption of the proposition, all fair equilibria in S_m^w are equivalent, we have that $N_m^w(f) = N_m^w(f')$, for all $f, f' \in \mathcal{F}_m$. Thus to prove the proposition it suffices to consider the following three cases.

Case 1: $N_m^w(c) > N_m^w(f)$**, for all** $f \in \mathcal{F}_m$ In this case we have that $N_l^h(c) \leq N_l^h(f) - 1$ for at least one $l \in \mathcal{K}_m$. First we prove, that without loss of generality, one can assume that the bottleneck agents for configuration c are connected to a hotspot. Indeed, we have:

$$U^h\left(N_l^h(c)+1\right) \geq U^h\left(N_l^h(f)\right),$$

and

$$U^w\left(N_m^w(c)\right) \leq U^w\left(N_m^w(f)+1\right),$$

by Assumption 2.3 on utilities. Since in equilibrium f we must have $U^h(N_l^h(f)) \geq U^w(N_m^w(f)+1)$ we arrive at:

$$U^h\left(N_l^h(c)+1\right) \geq U^h\left(N_l^h(f)\right) \geq U^w(N_m^w(f)+1) \geq U^w\left(N_m^w(c)\right).$$

Hence, $U^h\left(N_l^h(c)+1\right) \geq U^w\left(N_m^w(c)\right)$ and thus the utility of the bottleneck agent stays the same or improves when an agent is switched from the WAN AP w_m to hotspot h_l.

Thus if c is maximizing the bottleneck among all configurations of agents choices, the bottleneck agents could be assumed to be connected to a hotspot. However, consider $l = \arg\max_{k \in \mathcal{K}_m} N_k^h(c)$. Then any agent connected to the hotspot h_l is the bottleneck for configuration c. Thus, since no agent connected to the WAN is the bottleneck for c, we have $U^h(N_l^h(c)) < U^w(N_m^w(c))$. Then we have the following chain of inequalities:

$$U^h(N_l^h(c)) \;<\; U^w(N_m^w(c)) \underset{(a)}{\leq} U^w(N_m^w(f)+1) \underset{(b)}{\leq} U^h(N_l^h(f)),$$

where inequality (a) follows from the assumption of Case 1, and inequality (b) – from the fact that f is an equilibrium. Thus $U^h(N_l^h(c)) < U^h(N_l^h(f))$ which means that $N_l^h(c) \geq N_l^h(f) +$

1. Since f is a fair configuration, we have $\max_{k\in\mathcal{K}_m} N_k^h(f) \leq N_l^h(f)+1$. But then, $N_l^h(c) \geq \max_{k\in\mathcal{K}_m} N_k^h(f)$, and hence $U_m^{min}(c) \leq U_m^{min}(f)$.

Case 2: $N_m^w(c) < N_m^w(f)$**, for all** $f \in \mathcal{F}_m$ We first prove that no agents connected to the WAN can be the bottleneck for configuration c. Indeed, by assumption of this paragraph, we have that there exists at least one $l \in \mathcal{K}_m$ such that $N_l^h(c) \geq N_l^h(f)+1$. Now assume that the agents connected to the WAN are the bottleneck for configuration c, hence $U^w(N_m^w(c)) \leq U^h(N_k^h(c))$, for all $k \in \mathcal{K}_m$. Then we have the following chain of inequalities:

$$U^w(N_m^w(f)) \;<\; U^w(N_m^w(c)) \;\leq\; U^w(N_l^h(c)) \;\leq\; U^h(N_l^h(f)+1)\,.$$

Hence $U^w(N_m^w(f)) < U^h(N_l^h(f)+1)$ which contradicts the fact that the agents connected to h_l in configuration f are in equilibrium. This shows that no agent connected to the WAN could be the bottleneck for the configuration c.

It follows that the agents within the hotspot h_n, such that $n = \arg\max_{k\in\mathcal{K}_m} N_k^h(c)$ are the bottleneck. Since there exists l such that $N_l^h(c) \geq N_l^h(f)+1$, we have that $N_n^h(c) \geq \max_{k\in\mathcal{K}_m} N_n^h(f)$, by the fair structure of f. This yields that $U_m^{min}(c) \leq U_m^{min}(f)$, which we claimed to show.

Case 3: $N_m^w(c) = N_m^w(f)$**, for all** $f \in \mathcal{F}_m$ First, we show again that no agent connected to the WAN could be the bottleneck for configuration c. Indeed, since $\mathbf{N}_m(f) \not\sim \mathbf{N}(c)$ we have that, by fair structure of f, there exists at least one $l \in \mathcal{K}_m$ such that $N_l^h(c) \geq N_l^h(f)+1$. Assuming that the agents connected to the WAN are the bottleneck in configuration c, we have the following chain:

$$U^w(N_m^w(f)) = U^w(N_m^w(c)) \leq U^h(N_l^h(c)) \leq U^h(N_l^h(f)+1)\,.$$

Thus, $U^w(N_m^w(f)) \leq U^h(N_l^h(f)+1)$ indicating that f could not be an equilibrium configuration. This contradiction shows that the bottleneck agents for configuration c must be connected to hotspots. It is easy to see that $\max_{k\in\mathcal{K}_m} N_k^h(c) \geq \max_{k\in\mathcal{K}_m} N_k^h(f)$ which yields $U_m^{min}(c) \leq U_m^{min}(f)$.

Chapter 3

Shares of subscribers for competing providers

3.1 Introduction

In this chapter we will describe how to compute the fractions of agents that would be connected to WAN APs and hotspots in equilibrium. These fractions are useful as, for example, a metric for assessing "competitiveness" of one provider versus another. Note that the model that we built could readily be used to obtain the subscriber shares via simulation. However, as there are many elements involved in the model, (e.g. various capacities at access points, densities of access points and subscribers, geometric characteristics of coverage zones, etc.) the simulations might not yield a plausible enough demonstration of a possibly complex role that those parameters might play. Thus the objective of this section is to obtain *analytical* results that could be used to estimate subscriber shares and sensitivities of these to system parameters without simulation.

We will assume[1] that $\Pi^{w,\alpha}$ is a deterministic process such that the Voronoi cells associated with each WAN AP are geometrically similar and have the same area α. We further let the processes Π^h and Π^a be independent stationary Poisson processes with densities λ^h and λ^a respectively. Throughout the chapter we will also assume that the service zones of WAN and WLAN APs are defined as described in Section 2.3.2.

The non-uniqueness of equilibria poses certain difficulties in analyzing the model for arbitrary utilities, densities and cell sizes. Note that in practice, the sizes of WAN service zones typically would exceed that of hotspots[2]. Thus, to simplify our analysis we will study a system where the coverage area of WAN service zones, denoted α, is large enough to ensure each WAN service zone contains a large number of users and hotspots. Intuitively, one might expect that when the WAN service zones grow in area, the set of different equilibria becomes tighter, i.e.,

[1] This is, perhaps, not a bad assumption since WAN network would be carefully designed and optimized

[2] See e.g. [63] for a nice comparison of WiFi vs. 3G technologies.

a type of the Law of Large Numbers making the system more amenable to analysis. In the next section we demonstrate that this intuition is indeed correct.

3.2 Setup for asymptotic analysis

We consider a sequence of deterministic point processes $\{\Pi^{w,\alpha}\}$ indexed by $\alpha \in \mathbb{R}^+, \alpha \uparrow \infty$, where each represents the spatial locations of WAN APs which are increasingly spread out. In particular, we suppose that the area of the Voronoi cell associated with any point $w_m^\alpha \in \pi_\alpha^w$ is equal to α, and let α grow. Let us also assume that for each $\alpha > 0$, $\pi^{w,\alpha}$ contains a point w_0^α at the origin.

In what follows we will consider the service zones of WAN AP w_0^α and we will use the same notation as before to refer to the number of agents and hotspots falling within the service zone of the WAN AP $w_0^\alpha \in \pi^{w,\alpha}$, but indicate the dependence on the area α via the corresponding superscript. Thus, for example we will write H_0^α to indicate the number of hotspots that fall within the service zone $S_0^{w,\alpha}$ of the WAN AP w_0^α. In addition, we use $\mathbb{E}_0^h[A_0]$ to denote the expectation of the quantity A_k associated with a typical hotspot h_k (see, e.g. [64]), i.e. the expectation with respect to the Palm probability.

For fixed λ^h and λ^a the service area of each WAN AP will have to support a larger (roughly linear in α) number of users as α grows. Therefore, we will assume that the WAN resources also scale with α. This leads to a scaling requirement on the utility function associated with connecting to the WAN. Let $U^{w,\alpha}(\cdot)$ denote the utility function associated with connecting to the WAN when the area of a Voronoi cell of any WAN AP is α, and assume that $U^{w,\alpha}(\cdot)$ satisfies Assumption 2.3 for utility functions. Define $J^\alpha(N) = (U^{w,\alpha})^{-1} \circ U^h(\cdot)$ (where $U^h(\cdot)$ is independent of α) and assume the following:

Assumption 3.1. *The scaling of $J^\alpha(N)$ with α is such that:*

1. $J^\alpha(N) = \alpha j(N)$ *for any* $N \in \mathbb{N}$,
2. $\lim_{N\to\infty} j(N) = \infty$,
3. *There exists $\bar{N}$, such that* $j^{-1}\Big((N+1)/\alpha\Big) - j^{-1}\Big(N/\alpha\Big) < 1$, *for all* $N \geq \bar{N}$ *and each* $\alpha \geq 1$.
4. *For any integer* $K \geq 2$, $u(K) \neq j(K), j(K-1)$, *where*

$$u(K) = \lambda^a e^{-\lambda^h \pi d^2} + \lambda^h \mathbb{E}_0^h\left[(M_0^h - K + 1)\mathbf{1}_{\{M_0^h \geq K\}}\right]. \tag{3.1}$$

The interpretation of these assumptions is as follows. Condition 1 means that the resources of WAN APs scale linearly in the area α of their service zones. For example, we might

have $U^h(N) = \frac{B^h}{N}$ and $U^{w,\alpha}(N) = \frac{\alpha B^w}{N}$, in which case $J^\alpha(N) = \frac{\alpha B^w}{B^h}N$. The second condition follows if, as more agents connect to a resource, the utility of those agents is strictly decreasing to zero. The third condition allows us to use Proposition 2.6 to argue that the equilibrium in $S_m^{w,\alpha}$ is unique with probability approaching 1 as $\alpha \to \infty$. Finally, the last condition is technical, and satisfied for the cases of interest.

The asymptotics of this system are summarized in Theorem 3.1 and we give the details of the proof in Section 3.4. Here when we say that an event E^α happens with high probability (w.h.p.) we mean that $\lim_{\alpha\to\infty} \mathbb{P}(E^\alpha) = 1$.

Theorem 3.1. *Consider any realization of the Poisson point processes Π^a and Π^h and the sequence of deterministic processes $\{\Pi^{w,\alpha}\}$ with Voronoi cells of area α and each with a typical cell $S_0^{w,\alpha}$ centered at the origin coinciding with w_0^α. Under the scaling Assumption 3.1 we have:*

1. *The equilibrium f_0^α in $S_0^{w,\alpha}$ is unique and fair w.h.p.*
2. *The largest number of agents connected to each hotspot in this equilibrium, $N_{\max}^{h,\alpha}(f_0^\alpha) = \max_{k\in\mathcal{K}_0^\alpha} N_k^h(f_0^\alpha)$ has a limit:*

$$\lim_{\alpha\to\infty} N_{\max}^{h,\alpha}(f_0^\alpha) = N_{max}^{h,\infty},$$

   for some integer $N_{max}^{h,\infty} \geq 0$.
3. *We have that $N_{\max}^{h,\infty} > 0$ if and only if $j(1) \leq \lambda^a$ in which case it is given by the largest integer solution for $K \geq 1$ of the inequality*

$$u(K) \geq j(K), \tag{3.2}$$

   where $u(K)$ is given by (3.1).

We must note that item 3 in Theorem 3.1 requires certain properties of Poisson point processes. Items 1 and 2 actually will hold for scenarios where the point processes that describe APs' and users' locations are any stationary and ergodic (thus not necessarily Poisson), with a restriction that the variances of the number of APs and users within a given compact set on the plane asymptotically grow slower than some multiple of the corresponding means. The intuition is that for such processes, the standard deviation associated with the number of users located within a WAN service zone is asymptotically negligible in comparison to the mean.

The basic idea of the proof of Theorem 3.1 is to leverage the analogs of the Law of Large Numbers for functionals on random sets, e.g. Voronoi cells, which have distributions dependent on realizations of point processes. We also show that fluctuations from averages for the quantities of interest do not grow "too fast" as the area of the WAN service zones grows.

This allows us to express the position of the asymptotic "cutoff" $N_{max}^{h,\infty}$, in terms of averages of functionals of the realizations of Π^h and Π^a.

Based on Theorem 3.1 the analysis of competition when the WAN cell sizes are "large" reduces to comparing the number $N_{max}^{h,\infty} = K^*$ to the average number of agents falling within the service zone of a typical hotspot. In particular, if

$$K^* \gg \mathbb{E}_0^h\left[M_0^h\right] = \frac{\lambda^a(1-e^{-\lambda^h \pi d^2})}{\lambda^h}, \tag{3.3}$$

then hotspots retain most of the agents that fall within their service zones in equilibrium. We classify this case as hotspots effectively competing with the WAN. On the other hand if

$$K^* \ll \frac{\lambda^a(1-e^{-\lambda^h \pi d^2})}{\lambda^h}, \tag{3.4}$$

the hotspots yield most of their agents to the WAN APs in equilibrium. In this case we say that hotspots are not competitive with respect to the WAN. Using Theorem 3.1, we can suggest the following heuristic approach to estimate the value of N_{max}^h. In general one has to solve for $K \geq 0$ the equation:

$$U^w\left(\lambda^a |V| e^{-\lambda^h \pi |d|^2} + \lambda^h \mathbb{E}_0^h\left[P_0(K)\right]\right) = U^h(K), \tag{3.5}$$

where $P_0(K) = (M_0^h - K + 1)\mathbf{1}_{\{M_0^h \geq K\}}$. Note that since the left side of (3.5) is monotonically increasing in K and the right – monotonically decreasing, the solution either does not exist ($K^* = 0$) or is unique, when it exists. Unfortunately, there is no closed form expression for the term $\mathbb{E}_0^h[P_0(K)]$ and hence simulation has to be used to estimate it. However, to test if hotspots are not competitive with respect to the WAN one could use the following simple criterion. Clearly, (3.4) holds if the solution to:

$$U^w(\lambda^a \alpha - \lambda^h K \alpha) = U^h(K), \tag{3.6}$$

falls much below the value $\lambda^a/\lambda^h(1-e^{-\lambda^h \pi d^2})$. Note that this allows for a simple intuitive interpretation. The number of agents and hotspots occupying WAN service zone tends to $\lambda^a\alpha$ and $\lambda^h\alpha$ respectively when α is large. The number of agents connected in equilibrium to hotspots tends to $\lambda^h \alpha K$, whenever $K \ll \lambda^a \mathbb{E}|S_k^h|$, since then we can assume that each hotspot has exactly K agents connected to its AP in equilibrium. Thus the number of agents connected to the WAN AP must tend to:

$$\lambda^a \alpha - \lambda^h K \alpha,$$

once the size of the WAN service zone gets large enough. Thus, (3.6) follows by equating the utility of agents that are connected to the WAN AP and utility of the ones that are connected to hotspots.

3.3 Verification via simulation

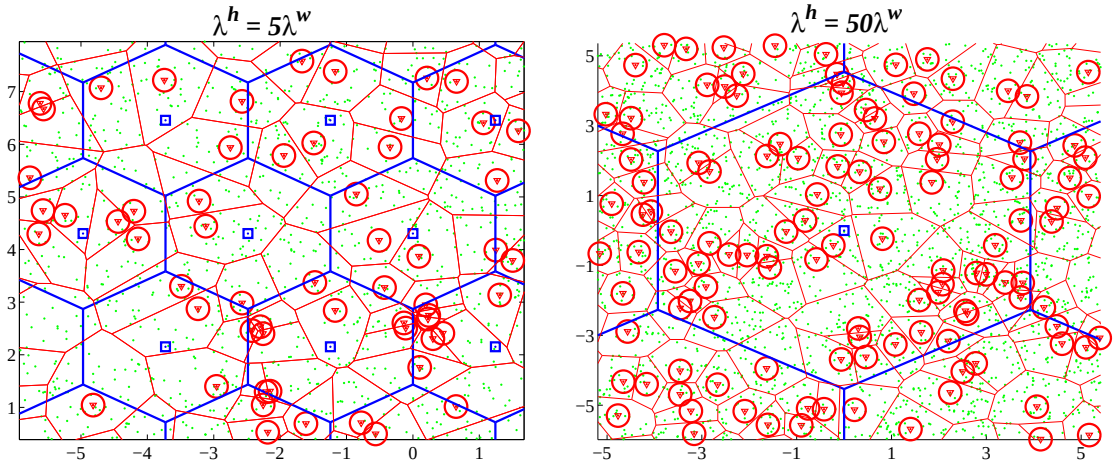


Figure 3.1: Representative geometries for different scaling factors α. The locations of users (green dots) and hotspots (red circles) are given by two independent Poisson processes of fixed densities. The service areas for WAN APs (blue hexagons) grow in area linearly in α. The figure on the left corresponds to $\alpha = 5$, and the one on the right – to $\alpha = 50$.

To support the asymptotics in Theorem 3.1 we will construct a simple simulation scenario. Let us consider the setup of Example 2.3, where utilities of agents connected to the WAN and hotspots are solely congestion-dependent. We let $U_j^{w,\alpha}(N) = \frac{\alpha B^w}{N}$, for $\alpha \geq 1$, and $U^h(N) = \frac{B^h}{N}$. We let $\Pi^{w,\alpha}$ to be a deterministic process that puts WAN APs on the plane in a honeycomb pattern, with density inversely proportional to the scaling factor α. At the same time, the locations of users and hotspots are given by two independent spatial Poisson processes with some fixed densities, independent of α. We use the construction of WAN and hotspot service zones described in Section 2.3.2. The representative geometries for two different scaling factors are shown in Figure 3.3, and the summary of different parameters that we are using is given in Table 3.1

Parameter	Notation	Value
Agents' density	λ^a	20 km^{-2}
Hotspots' density	λ^h	1 km^{-2}
WAN APs' density	λ^w/α	$0.81/\alpha$ km^{-2}
Hotspot coverage radius	d	500 m
Hotspots' bandwidth	B^h	1.5 Mbs
WAN APs' bandwidth	αB^w	3α Mbs

Table 3.1: Simulation parameters

Figure 3.2 exhibits the results of a series of experiments in which we have simulated

the area corresponding to 10×10 WAN service zones each with increasing coverage area. The figure shows the distribution of equilibrium "cutoff level" $N_{max}^{h,\alpha}$, i.e., the distribution of the largest number of agents within any given WAN service zone that connect to a hotspot in equilibrium. We note that the width of the distribution shrinks when we increase the scaling α and eventually all mass of the distribution concentrates at point $K^* = N_{max}^{h,\infty} = 7$. This behavior of the cutoffs' distribution function is in perfect agreement with Theorem 3.1, and we verified that the asymptotic value of the cutoff indeed agrees with the one predicted item 3 in Theorem 3.1.

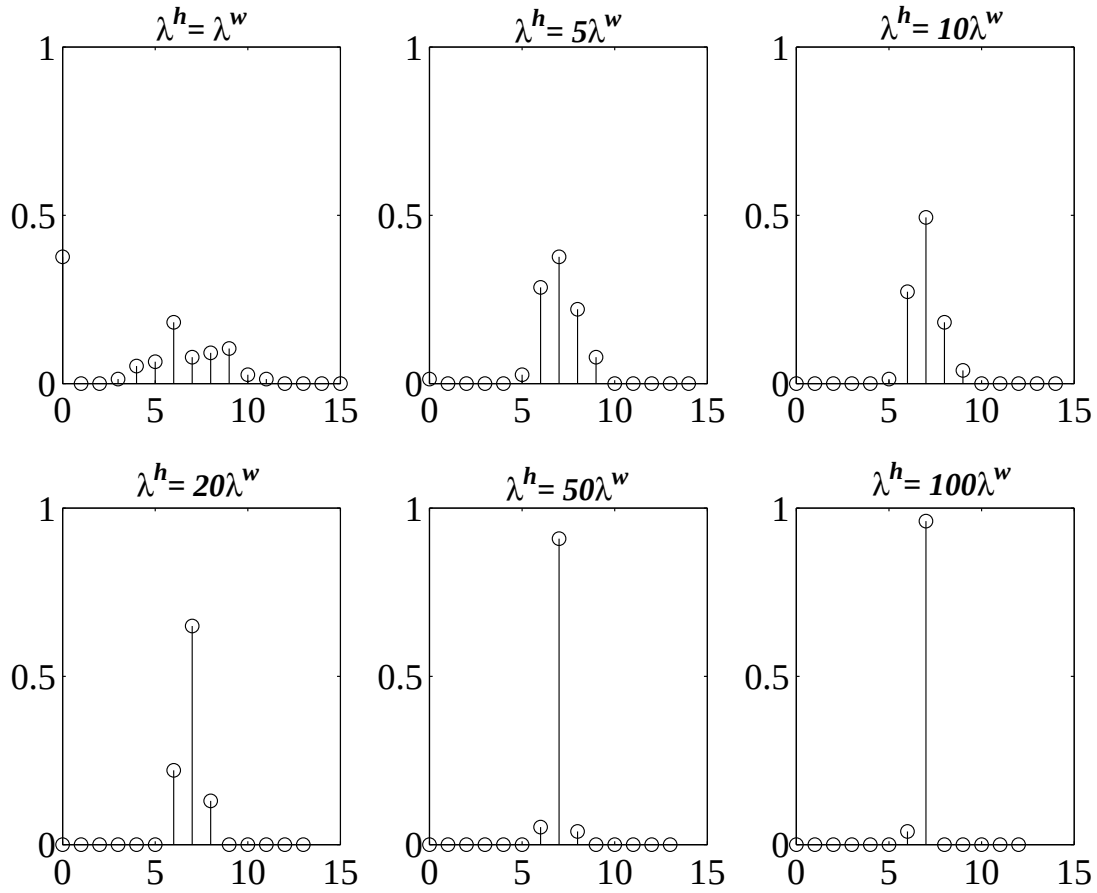


Figure 3.2: The evolution of the sample distribution function for the "cutoff level" N_{max}^h for different scaling factors α.

3.4 Proof of Theorem 3.1

Prior to giving a proof of Proposition 3.1 we provide several technical lemmas.

Lemma 3.1. *For any realization of the Poisson processes Π^a and Π^h consider a service zone*

associated with the WAN AP $w_0^{\alpha} \in \pi^{w,\alpha}$. *Let*

$$L_k(K) = (M_k^h - K)\mathbf{1}_{\left\{M_k^h \geq K\right\}}, \quad P_k(K) = (M_k^h - K + 1)\mathbf{1}_{\left\{M_k^h \geq K\right\}}. \tag{3.7}$$

For any $m \in \mathbb{N}$ *we have the following a.s. limits:*

$$\lim_{\alpha\to\infty} \frac{H_0^{\alpha}}{\alpha} = \lambda^h, \quad \lim_{\alpha\to\infty} \frac{M_0^{w,\alpha}}{\alpha} = \lambda^a, \tag{3.8}$$

$$\lim_{\alpha\to\infty} \frac{\sum_{k\in\mathcal{K}_0^{\alpha}} L_k(K)}{H_0^{\alpha}} = \mathbb{E}_0^h\left[L_0(K)\right], \quad \lim_{\alpha\to\infty} \frac{\sum_{k\in\mathcal{K}_0^{\alpha}} P_k(K)}{H_0^{\alpha}} = \mathbb{E}_0^h\left[P_0(K)\right], \tag{3.9}$$

$$\lim_{\alpha\to\infty} \frac{M_{C_0}^{w,\alpha}}{\alpha} = \lambda^a(1 - e^{-\lambda^h \pi |d|^2}), \qquad \lim_{\alpha\to\infty} \frac{M_{\bar{C}_0}^{w,\alpha}}{\alpha} = \lambda^a e^{-\lambda^h \pi |d|^2}, \tag{3.10}$$

$$\lim_{\alpha\to\infty} \mathbb{P}\left(\exists h_k \in S_0^{w,\alpha} : M_k^h \geq K\right) = 1, \ \forall K \geq 0. \tag{3.11}$$

Proof. The limits (3.8) follow by ergodicity [53] of the process π^a and π^h. One needs only to note that the ratio $\alpha/|S_0^{w,\alpha}|$ converges to 1 as $\alpha \to \infty$ since d (the radius of hotspot coverage) is bounded.

Consider now the limits (3.9). Note that for each k, and any fixed K, both $L_k(K)$ and $P_k(K)$ are functionals of the realization of processes Π^h and Π^a within some a.s. bounded region (Voronoi "flower" [29] associated with the Voronoi cell V_k^h). Thus $L_k(K)$ and $P_k(K)$ are "local statistics" as defined in [29], and thus one can use Theorem 3.1 therein to obtain these limits.

Now consider the limits (3.10). By (3.8) and (3.9) and noting that:

$$\sum_{k\in\mathcal{K}_0^{\alpha}} M_k^h = \sum_{k\in\mathcal{K}_0^{\alpha}} L_k(K)|_{K=0},$$

we have:

$$\lim_{\alpha\to\infty} \frac{\sum_{k\in\mathcal{K}_0^{\alpha}} M_k^h}{\alpha} = \lambda^h \mathbb{E}_0^h\left[M_0^h\right].$$

Evaluating this expectation, we get:

$$\mathbb{E}_0^h\left[M_0^h\right] \quad = \mathbb{E}_0^h\left[\sum_{a_i\in\Pi^a(V_0^h)} \mathbf{1}_{\left\{a_i\in V_0^h\right\}} \mathbf{1}_{\{|a_i|\leq d\}}\right] = \mathbb{E}_0^h\left[\sum_{a_i\in\Pi^a} \mathbf{1}_{\left\{\Pi^h(B(a_i,|a_i|))=\emptyset\right\}} \mathbf{1}_{\{|a_i|\leq d\}}\right],$$

where the second equality uses the fact that if $a_i \in V_0^h$ then there can be no other point of Π^h

within the ball of radius $|a_i|$ centered at a_i. Now by independence of Π^h and Π^a and also using Campbell's formula and Slyvnyak's theorem (see e.g. [65]) we get:

$$\begin{aligned}\mathbb{E}_0^h\left[M_0^h\right] &= \mathbb{E}_0^h\left[\int_{x\in B(0,d)} \mathbf{1}_{\left\{\Pi^h(B(x,|x|))=\emptyset\right\}}\lambda^a dx\right] = \int_{x\in B(0,d)} e^{-\lambda^h\pi|x|^2}\lambda^a dx \\ &= \frac{\lambda^a}{\lambda^h}(1-e^{-\lambda^h\pi|d|^2}),\end{aligned}$$

from which the first limit in (3.10) follows. The second limit in (3.10) follows by taking into account the limit (3.9) and the first limit in (3.10).

Finally, to obtain the limit (3.11), we apply the Strong Law of Large Numbers to the sum of random variables $Z_k \triangleq \mathbf{1}_{\left\{M_k^h>K\right\}}$ to obtain:

$$\lim_{\alpha\to\infty}\frac{1}{H_0^\alpha}\sum_{k\in\mathcal{K}_0^\alpha}\mathbf{1}_{\left\{M_k^h>K\right\}} = \lim_{\alpha\to\infty}\frac{1}{H_0^\alpha}\sum_{k\in\mathcal{K}_0^\alpha} Z_k = \mathbb{P}(M_k^h>K) > 0 \quad a.s.. \tag{3.12}$$

Here we used the fact that the variables Z_k are i.i.d., since they depend on the number of points of homogeneous Poisson process sampled on disjoint sets S_k^h. Thus, at least one term in the sum in (3.12) is nonzero, for sufficiently large α, which proves the limit (3.11). □

Lemma 3.2. *Let Δ_i^α where $i = 1,2,3,4$ be defined as follows:*

$$\Delta_1^\alpha = M_0^{w,\alpha}, \quad \Delta_2^\alpha(K) = \sum_{k\in\mathcal{K}_0^\alpha} L_k(K), \quad \Delta_3^\alpha(K) = \sum_{k\in\mathcal{K}_0^\alpha} P_k(K), \quad \Delta_4^\alpha = M_{\tilde{C}_0}^{w,\alpha}.$$

Then for each i, $1\le i\le 4$ and any $C>0$ we have:

$$\lim_{\alpha\to\infty}\mathbb{P}\left[|\Delta_i^\alpha - \mathbb{E}[\Delta_i^\alpha]| > C\sqrt{\alpha\log\alpha}\right] = 0. \tag{3.13}$$

Proof. To prove the lemma we will use Chebyshev's inequality:

$$\mathbb{P}\left[|\Delta_i^\alpha - \mathbb{E}[\Delta_i^\alpha]| > C\sqrt{\alpha\log\alpha}\right] \le \frac{\mathbf{var}[\Delta_i^\alpha]}{C^2\,\alpha\log\alpha}.$$

First we show that for $1\le i\le 4$:

$$\mathbf{var}[\Delta_i^\alpha] = O(\alpha). \tag{3.14}$$

Indeed, $\Delta_1^\alpha = M_0^{w,\alpha}$ is just a Poisson random variable with average that scales linearly in α. Hence (3.14) is satisfied for $i=1$. To obtain the bound on the variances of Δ_2^α and Δ_3^α we use Lemma 1 in [29], which yields:

$$\mathbf{var}[\Delta_2^\alpha(K)] = O(\alpha), \quad \mathbf{var}[\Delta_3^\alpha(K)] = O(\alpha).$$

Finally for the variance of $M^{w,\alpha}_{\tilde{C}_0}$ observe that

$$M^{w,\alpha}_{\tilde{C}_0} = M^{w,\alpha}_0 - \Delta^\alpha_2(0)\,.$$

Since the variances of both terms on the right are $O(\alpha)$ we get:

$$\mathbf{var}\,[\Delta^\alpha_4] = O(\alpha)\,.$$

Now using Chebychev's inequality and (3.14) we obtain, for any $C > 0$,

$$\mathbb{P}\left(|\Delta^\alpha_i - \mathbb{E}\,[\Delta^\alpha_i]| > C\sqrt{\alpha\log\alpha}\right) = \frac{O(\alpha)}{\Theta(\alpha\log\alpha)} \to 0\,,\ \text{when } \alpha\to\infty\,.$$

□

Lemma 3.3. *Under the scaling Assumption 3.1, the equilibrium f^α_0 in $S^{w,\alpha}_0$ is unique and fair w.h.p..*

Proof. Using Lemma 3.2 we have that, eventually, $M^{w,\alpha}_{\tilde{C}_0} \geq \bar{N}$ a.s. as $\alpha\to\infty$. Taking into account Assumption 3.1, the conditions of Proposition 2.6 hold w.h.p. Using Proposition 2.6 yields the statement of the lemma. □

Lemma 3.4. *For any equilibrium configuration f^α_0 in $S^{w,\alpha}_0$ we have that:*

$$\max_{k\in\mathcal{K}^\alpha_0} N^h_k(f^\alpha_0) < \max_{k\in\mathcal{K}^\alpha_0} M^h_k\,,\ \textit{w.h.p.} \tag{3.15}$$

Proof. Note that (3.15) has a strict inequality. Thus (3.15) implies that the largest number of agents connected to any hotspot within S^w_0 in equilibrium f^α_0 is strictly less than the maximum number of agents in any one of the hotspots – at least asymptotically. We prove the lemma by contradiction. Suppose that there exists a sequence $\xi^\varepsilon = \{\alpha_n > 0 | \lim_{n\to\infty}\alpha_n = \infty\}$ with the following property. For any $\alpha\in\xi^\varepsilon$, f^α_0 is such that for some $l^\alpha \in \mathcal{K}^\alpha_0$ we have $N^h_{l^\alpha}(f^\alpha_0) = \max_{k\in\mathcal{K}^\alpha_0} M^h_k$ with probability greater than ε. Then, for any $\alpha\in\xi^\varepsilon$:

$$J^\alpha(M^h_{l^\alpha}) \leq N^{w,\alpha}_0(f^\alpha_0)\,, \tag{3.16}$$

since no agent desires to switch to the WAN AP w_0 from the hotspot h_{l^α}. Now, note that f^α_0 is fair w.h.p, by Lemma 3.3 and thus:

$$M^h_k - 1 \leq N^h_k(f^\alpha_0) \leq M^h_k\,,$$

where we took into account that there are no $k\in\mathcal{K}^\alpha_0$ such that $M^h_k > M^h_{l^\alpha}$. This yields, that at

most one agent within each hotspot h_k, for $k \in \mathcal{K}_0^\alpha$ selects the WAN, thus

$$N_0^{w,\alpha}(f_0^\alpha) \leq M_{\tilde{C}_0}^{w,\alpha} + H^m\,, \tag{3.17}$$

Now, using Assumption 3.1 and Lemma 3.1, the inequalities (3.16) and (3.17) imply:

$$j\left(M_{l^\alpha}^h\right) \leq \lambda^a e^{-\lambda^h \pi d^2} + \lambda^h\,. \tag{3.18}$$

Taking into account that by Lemma 3.1 and Assumption 3.1:

$$\liminf_{\alpha\to\infty} \max_{k\in\mathcal{K}_0^\alpha} M_k^h = \infty, \qquad \lim_{N\to\infty} j(N) = \infty,\ a.s.$$

we find that the inequality (3.18) is violated with probability tending to 1 as $\alpha \to \infty$. Thus, ξ^ε can not exist for any $\varepsilon > 0$, which proves the lemma. □

Lemma 3.5. *Consider a configuration f_0^α for service zone $S_0^{w,\alpha}$ and let $N_{max}^{h,\alpha}(f_0^\alpha) = \max_{k\in\mathcal{K}_m} N_k^h(f_0^\alpha)$. For any $\alpha > 0$, a necessary and sufficient condition for f_0^α to be an equilibrium w.h.p. is that f_0^α is a fair configuration that obeys either of the following:*

$$N_{max}^{h,\alpha}(f_0^\alpha) = 0\,, \quad J^\alpha(1) > M_0^{w,\alpha}\,, \tag{3.19}$$

$$N_{max}^{h,\alpha}(f_0^\alpha) \geq 1\,, \quad J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha)\right) - 1 \leq N_0^{w,\alpha}(f_0^\alpha) < J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha) + 1\right) \tag{3.20}$$

where $N_k^h(f_0^\alpha) = N_{max}^{h,\alpha}(f_0^\alpha)$ for all $k \in \mathcal{K}_0^\alpha$, such that $M_k^h \geq N_{max}^{h,\alpha}(f_0^\alpha)$, or:

$$N_{max}^{h,\alpha}(f_0^\alpha) \geq 1\,, \quad J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha)\right) - 1 \leq N_0^{w,\alpha}(f_0^\alpha) < J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha)\right)\,, \tag{3.21}$$

where $\exists k,l \in \mathcal{K}_0^\alpha$, such that $M_k^h, M_l^h \geq N_{max}^{h,\alpha}(f_0^\alpha)$, and $M_k^h = N_{max}^{h,\alpha}(f_0^\alpha)$, $M_l^h = N_{max}^{h,\alpha}(f_0^\alpha) - 1$.

Proof. We already proved in Lemma 3.3 that all equilibria in $S_0^{w,\alpha}$ have the same fair characterizations w.h.p. In case $N_{max}^{h,\alpha}(f_0^\alpha) = 0$ there are no agents connected to any hotspots in S_0^w. The necessary and sufficient condition for that, as follows from the inequality (2.5), is given by (3.19).

Consider the case $N_{max}^{h,\alpha}(f_0^\alpha) \geq 1$. First assume that for all $k \in \mathcal{K}_0^\alpha$, such that $N_{max}^{h,\alpha}(f_0^\alpha)$ we have that $N_k^h(f_0^\alpha) = N_{max}^{h,\alpha}(f_0^\alpha)$. By Lemma 3.4 we have $N_{max}^{h,\alpha}(f_0^\alpha) < \max_{k\in\mathcal{K}_0} M_k^h$, and thus we can use the equilibrium conditions (2.4) to obtain (3.20).

Now assume, instead, that there exist such $k,l \in \mathcal{K}_0^\alpha$, so that $M_k^h, M_l^h \geq N_{max}^{h,\alpha}(f_0^\alpha)$, and $M_k^h = N_{max}^{h,\alpha}(f_0^\alpha)$, $M_l^h = N_{max}^{h,\alpha}(f_0^\alpha) - 1$. For the hotspots having $N_{max}^{h,\alpha}(f_0^\alpha) - 1$ agents connected

to them in configuration f_0^α, via equilibrium conditions (2.4) we get:

$$J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha)-1\right)-1 \le N_0^{w,\alpha}(f_0^\alpha) < J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha)\right), \tag{3.22}$$

while for the hotspots having $N_{max}^{h,\alpha}(f_0^\alpha)$ agents connected to them:

$$J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha)\right)-1 \le N_0^{w,\alpha}(f_0^\alpha) < J^\alpha\left(N_{max}^{h,\alpha}(f_0^\alpha)+1\right), \tag{3.23}$$

Now, using monotonicity of $J^\alpha(\cdot)$, by combining (3.22) and (3.23) we get (3.21). □

Proof of Theorem 3.1 Let f_0^α denote an equilibrium configuration in the service zone $S_0^{w,\alpha}$ of the WAN AP $w_0^\alpha \in \pi_\alpha^w$. By Lemma 3.3 such configurations have equivalent and fair characterizations w.h.p, which gives Part 1 of the theorem. Let $N_{max}^{h,\alpha} = \max_{k\in\mathcal{K}_0^\alpha} N_k^h(f)$ where $f \in \mathcal{F}_0^\alpha$ is any fair equilibrium configuration. In what follows we will consider two cases that depend on whether the density of agents λ^a is less than the value $j(1)$. Our goal is to show that the $\lim_{\alpha\to\infty} N_{max}^{h,\alpha}(f_0^\alpha)$ exists.

Case 1: $\lambda^a < j(1)$ We will show that $\lambda^a < j(1)$ if and only if:

$$\lim_{\alpha\to\infty} N_{max}^{h,\alpha} = 0\,.$$

Indeed, the "only if" part follows from the condition (3.19) by dividing both sides by α and taking limits as $\alpha\to\infty$. Now using the limit (3.8) we obtain that $N_{max}^{h,\alpha} = 0$ w.h.p. implies $\lambda^a < j(1)$.

Next we prove that if $\lambda^a < j(1)$ then $N_{max}^{h,\alpha} = 0$ w.h.p. Indeed, by Lemma 3.1 we know that:

$$M_0^{w,\alpha} = \lambda^a\alpha + \varepsilon(\alpha)\,,$$

where $|\varepsilon(\alpha)| = O\left(\sqrt{\alpha\log\alpha}\right)$. But then, for sufficiently large α we have:

$$M_0^{w,\alpha} < J^\alpha(1)\,,$$

which, by Lemma 3.5 implies $N_{max}^{h,\alpha} = 0$ w.h.p.

Case 2: $\lambda^a \ge j(1)$ We first prove that $N_{max}^{h,\alpha}$ has a limit once $\alpha\to\infty$. Consider any sequence $\xi := \{\alpha_n | n\in\mathbb{N}\}$, where $\lim_{n\to\infty}\alpha_n = \infty$. We define the following disjoint subsequences of ξ:

$$\xi_1 = \left\{\alpha |\ \alpha\in\xi,\ 1 \le N_{max}^{h,\alpha} < \max_{k\in\mathcal{K}_0^\alpha} M_k^h \text{ and } \forall k\in\mathcal{K}_0^\alpha, \text{ s.t. } M_k^h \ge N_{max}^{h,\alpha},\ N_k^h(f^\alpha) = N_{max}^{h,\alpha}\right\}$$

$$\xi_2 = \Bigg\{ \alpha |\ \alpha \in \xi,\ 1 \le N_{max}^{h,\alpha} < \max_{k \in \mathcal{K}_0^\alpha} M_k^h\ ,\ \text{and}$$
$$\exists k,l \in \mathcal{K}_0^\alpha,\ \text{s.t. } M_k^h, M_l^h \ge N_{max}^{h,\alpha},\ \text{and } N_k^h(f^\alpha) = N_{max}^{h,\alpha},\ N_l^h(f^\alpha) = N_{max}^{h,\alpha} - 1 \Bigg\}$$

$$\xi_3 = \left\{ \alpha |\ \alpha \in \xi,\ 1 \le N_{max}^{h,\alpha} = \max_{k \in \mathcal{K}_0^\alpha} M_k^h \right\}$$

$$\xi_4 = \left\{ \alpha |\ \alpha \in \xi,\ N_{max}^{h,\alpha} = 0 \right\}$$

Clearly, $\xi = \bigcup_{i=1}^4 \xi_i$. However, by Lemma 3.4, the sequence ξ_3 is finite. Moreover, we have proved above that when $\lambda^a \ge j(1)$ the sequence ξ_4 is finite too. Thus, asymptotically, ξ consists only of the members of the sequences ξ_1 and ξ_2. Note that either ξ_1 or ξ_2 or both ξ_1 and ξ_2 have to be infinite, since ξ is infinite.

By definition of ξ_1 and ξ_2, we have, that $\max_{k \in \mathcal{K}_0^\alpha} M_k^h > N_{max}^{h,\alpha}$ when $\alpha \in \xi_1 \cup \xi_2$. Thus, if any of ξ_1 or ξ_2 is finite, to prove the statement of the theorem we have to show that $N_{max}^{h,\alpha}$ converges along the other infinite sequence. If both ξ_1 and ξ_2 are infinite, then we need to show that $N_{max}^{h,\alpha}$ is asymptotically the same along each subsequence, and in addition that:

$$\lim_{\alpha \in \xi_1,\ \alpha \to \infty} N_{max}^{h,\alpha} = \lim_{\alpha \in \xi_2,\ \alpha \to \infty} N_{max}^{h,\alpha}.$$

We will consider first the sequence ξ_1 and assume that it is infinite. We will show that:

$$\lim_{\alpha \in \xi_1,\ \alpha \to \infty} N_{max}^{h,\alpha} = K_1\,, \tag{3.24}$$

where K_1 is independent of α. We argue by contradiction. In particular, assume that there exist arbitrary large $\gamma, \delta \in \xi_1$ such that $N_{max}^{h,\gamma} \ne N_{max}^{h,\delta}$. Without loss of generality let $\gamma < \delta$, and consider the equilibrium conditions in $S_0^{w,\delta}$. By Lemma 3.5 we have that:

$$J^\delta\left(N_{max}^{h,\delta}\right) - 1 \le N_0^w(f_0^\delta) < J^\delta\left(N_{max}^{h,\delta} + 1\right),$$

Now multiplying these inequalities by γ/δ, and using Assumption 3.1, we obtain:

$$J^\gamma\left(N_{max}^{h,\delta}\right) - \gamma/\delta \le \gamma/\delta N_0^{w,\delta}(f_0^\delta) < J^\gamma\left(N_{max}^{h,\delta} + 1\right).$$

Note that by Lemma 3.1 we have:

$$\gamma/\delta N_0^{w,\delta}(f_0^\delta) = \gamma\left(\lambda^a e^{-\lambda^h \pi d^2} + \lambda^h \mathbb{E}_0^h\left[P_0^\delta(N_{max}^{h,\delta})\right]\right) + \varepsilon_1(\gamma,\delta)\,,$$

where, by Lemma 3.2, $|\varepsilon_1(\gamma,\delta)| = O\left(\gamma\sqrt{\log\delta/\delta}\right) = O\left(\sqrt{\gamma\log\gamma}\right)$. This yields:

$$J^\gamma\left(N_{max}^{h,\delta}\right) - 1 \leq \gamma\left(\lambda^a e^{-\lambda^h\pi d^2} + \lambda^h\mathbb{E}_0^h\left[P_0^\delta(N_{max}^{h,\delta})\right]\right) + \varepsilon_1(\gamma,\delta) < J^\gamma\left(N_{max}^{h,\delta}+1\right). \tag{3.25}$$

Now consider a fair configuration $\tilde{f}_0^\gamma$ for service zone $S_0^{w,\gamma}$, such that $\max_{k\in\mathcal{K}_0^\alpha} N_k^h(\tilde{f}_0^\gamma) = N_{max}^{h,\delta}$ and such that for all $k\in\mathcal{K}_0^\gamma$ for which $M_k^h \geq N_{max}^{h,\gamma}$ we have $N_k^h(\tilde{f}_0^\gamma) = N_{max}^{h,\delta}$. Clearly, in this case Lemma 3.1 and Lemma 3.2 yield:

$$N_0^w(\tilde{f}_0^\gamma) = \gamma\left(\lambda^a e^{-\lambda^h\pi d^2} + \lambda^h\mathbb{E}_0^h P_0^\delta\left(N_{max}^{h,\delta}\right)\right) + \varepsilon_2(\gamma),$$

where $|\varepsilon_2(\gamma)| = O\left(\sqrt{\gamma\log\gamma}\right)$. By Assumption 3.1 (item 4) we have, for all integer K:

$$\lambda^a e^{-\lambda^h\pi d^2} + \lambda^h\mathbb{E}_0^h P_0^\delta(K) \neq j(K),$$

which then translates the inequalities (3.25) into:

$$j\left(N_{max}^{h,\delta}\right) < \lambda^a e^{-\lambda^h\pi d^2} + \lambda^h\mathbb{E}_0^h P_0^\delta\left(N_{max}^{h,\delta}\right) < j\left(N_{max}^{h,\delta}+1\right). \tag{3.26}$$

Now note that $|\varepsilon_1(\gamma,\delta)+\varepsilon_2(\gamma)| = O\left(\sqrt{\gamma\log\gamma}\right) = o(\gamma)$. Hence, using (3.25), one gets

$$J^\gamma\left(N_{max}^{h,\delta}\right) - \gamma/\delta \leq N_0^{w,\gamma}(\tilde{f}_0^\gamma) < J^\gamma\left(N_{max}^{h,\delta}+1\right),$$

once $\gamma < \delta$ are selected large enough. By Lemma 3.5 we have that $\tilde{f}_0^\gamma$ is a fair equilibrium that is different from f_0^γ. By Lemma 3.3 this can not happen w.h.p. Thus we obtain that

$$\lim_{\alpha\in\xi_1,\alpha\to\infty} N_{max}^{h,\delta} = K_1$$

for some positive integer K_1.

Now, if ξ_2 is finite, we are done, since asymptotically ξ consists only of ξ_1 and we have already shown that along ξ_1 the value of $N_{max}^{h,\alpha}$ has a limit. Now we will prove that if ξ_2 is infinite, then the value of $N_{max}^{h,\alpha}$ along ξ_2 also converges to a limit. Take any $\gamma\in\xi_2$ then, by Lemma 3.5 we have, that

$$J^\alpha(N_{max}^{h,\gamma}) - 1 \leq N_0^{w,\gamma}(f^\gamma) < J^\alpha(N_{max}^{h,\gamma}).$$

Dividing these inequalities by γ, by Assumption 3.1, we have:

$$j(N_{max}^{h,\gamma}) - 1/\gamma \leq \frac{N_0^w(f^\gamma)}{\gamma} < j(N_{max}^{h,\gamma}). \tag{3.27}$$

We now show that $N_{max}^{h,\gamma} < K_0$ for some K_0 independent of γ. Indeed, otherwise, there exists a subsequence $\xi_5 \subset \xi_2$, with $\lim_{\gamma\to\infty,\gamma\in\xi_5} N_{max}^{h,\gamma} = \infty$. Now using Lemma 3.1, we have that

$$\limsup_{\gamma\to\infty,\gamma\in\xi_5} \frac{N_0^w(f^\gamma)}{\gamma} < \lambda^a e^{-\lambda^h \pi d^2} + \lambda^h \lim_{\gamma\to\infty,\gamma\in\xi_5} \mathbb{E}_0^h \left[P_0^\gamma \left(N_{max}^{h,\gamma} \right) \right] = \lambda^a e^{-\lambda^h \pi d^2}.$$

At the same time we have that

$$\lim_{\gamma\to\infty,\gamma\in\xi_5} j(N_{max}^{h,\gamma}) = \infty,$$

which means that the inequalities (3.27) could not be satisfied along the subsequence ξ_5. Thus we have a contradiction, and $\exists K_0$, such that $N_{max}^{h,\gamma} < K_0$.

Thus the set $\{N_{max}^{h,\gamma} | \gamma \in \xi_2\}$ is finite, hence if ξ_2 is infinite, at least some values from this set must realize infinitely often along ξ_2. Consider any value K which is achieved infinitely often along ξ_2, i.e. there exists a subsequence $\xi_6 \subset \xi_2$, with $\sup\{\gamma | \gamma \in \xi_6\} = \infty$ and for any $\gamma \in \xi_6$, $N_{max}^{h,\gamma} = K$. Note that $N_0^{w,\gamma}(f_0^\gamma)$ must satisfy:

$$M_{\bar{C}_0}^{w,\gamma} + \sum_{k\in\mathcal{K}_0^\gamma} (M_k^h - K)\mathbf{1}_{\{M_k^h \geq K\}} < N_0^w(f_0^\gamma) < M_{\bar{C}_0}^{w,\gamma} + \sum_{k\in\mathcal{K}_0^\gamma} (M_k^h - K + 1)\mathbf{1}_{\{M_k^h \geq K\}} \ w.h.p.,$$

since f_0^γ is asymptotically fair w.h.p. Dividing these inequalities by γ, and comparing to inequalities (3.27), one finds that K must satisfy:

$$l(K) \triangleq \lambda^a e^{\lambda^h \pi d^2} + \lambda^h \mathbb{E}_0^h [L_0(K)] < j(K) \leq \lambda^a e^{\lambda^h \pi d^2} + \lambda^h \mathbb{E}_0^h [P_0(K)] \triangleq u(K), \tag{3.28}$$

where we used Lemma 3.1. Note that $u(K) = l(K+1)$ and thus the intervals $(l(K), u(K)]$ are disjoint for different integer K. Moreover $\bigcup_{K=1}^\infty (l(K), u(K)] = (0, u(1)]$. Since $j(K)$ is increasing in K and $\lim_{K\to\infty} j(K) = \infty$, there exists exactly one integer solution to the inequalities (3.28), since we assumed

$$j(1) \leq \lambda^a,$$

and $u(1) = \lambda^a$. But then the value of $N_{max}^{h,\gamma}$ is asymptotically unique w.h.p., when $\gamma \in \xi_2$ and $\gamma \to \infty$.

We are left to show that if both ξ_1 and ξ_2 are infinite, then the asymptotic values K_1 and K_2 along ξ_1 and ξ_2 respectively satisfy $K_1 = K_2$. Observe that the condition (3.26) implies for K_1:

$$j(K_1) < u(K_1) < j(K_1 + 1).$$

Now, since K_2 is a unique integer solution to (3.28) we obtain that $K_2 = K_1$. Since $N_{max}^{h,\alpha}(f_0^\alpha) = K_1$ w.h.p. when $\alpha \in \xi_1$ and $N_{max}^{h,\alpha}(f_0^\alpha) = K_2 = K_1$ w.h.p when $\alpha \in \xi_2$, we obtain Part 2 of the theorem. Lastly, Part 3 of the theorem follows from the above analysis.

3.5 Summary of Chapters 2 and 3

In summary, at this point we have developed a geometric model for a system where subscribers with dual mode devices and select among two noninterfering wireless service providers – a WAN provider and a second provider (or aggregator) of LAN hotspots. Our model is of interest in that, on the one hand, it captures wireless providers using technologies that might have different capacity and coverage, and on the other hand it captures the role of subscribers decision-making mechanisms in determining the eventual equilibrium. Assuming that each subscriber makes greedy decisions, based on comparing two utilities, at random times, we show that an equilibrium configuration would eventually be reached. Further we have characterized such equilibria and shown that they are likely to be close to an equilibrium, corresponding to slicing the excess loads on hotspots, and "shifting" only those to the WAN. In an effort to get numerical estimates for the level at which this slicing occurs, we developed an asymptotic result for the case where WAN service areas are large, which would permit an evaluation of this setting.

The results of Chapter 3 can be viewed from different perspectives. On the one hand they permit an evaluation of the competitiveness of the two providers to attract subscribers in their service areas. We admit however, that such evaluation is contingent on knowing (or ability to model appropriately) users' utility functions. However, more importantly, and as we will see in the next chapter, our results permit a study of how to design decision making mechanisms, i.e., appropriate utility functions, to *realize* equilibria that may be desirable for the overall system. The highlight of Chapters 2,3 is characterization of such equilibria, and this would permit further consideration of the performance and network design implications of wireless systems where users are capable to switch among multiple providers, depending on the key parameters of the system.

Chapter 4

Design of cooperative multi-provider wireless networks under the "loosely coupled" option

4.1 Introduction

As WAN and WLAN technologies posses complementary features, cooperation scenarios between WAN and hotspots' providers could be of high practical value and can be realized in a number of different ways. For example, a WAN provider could be interested in augmenting its service via high bandwidth hotspots' access points in urban areas seeing a high steady load and thus seek an agreement with a hotspot provider which will let the subscribers of WAN service use hotspots within these locations. Similarly, a hotspot provider might be interested in upgrading its service which is only available locally at a limited set of locations, to a ubiquitous "always on" wireless access service, and thus seek an agreement which would oblige a WAN provider to serve hotspot provider's subscribers at locations not covered by hotspots. Finally, WAN and hotspot providers might merge together to become a single entity which provides a service possessing the complementary strengths of both WAN and hotspots' technologies.

In all of the above examples the resulting heterogeneous network can be designed according to several architectural options, which can be referred to as "loosely coupled" and "tightly coupled" internetworking solutions [11]. According to the former option the WAN and hotspots networks are kept as separate entities, but the subscriber databases for both providers are augmented so that the WAN and hotspots are able to handle the requests associated with the common subscribers of the WAN and the hotspots networks. This option does not assume any coordination between the access points or any additional feedback that the access points might provide to users to facilitate their decision-making (selection of services, when several options are available). As such, the simplicity of this design option implies that control on performance

of such networks will be quite limited.

The other design option is especially suitable in a scenario where the WAN and hotspots belong to the same service provider (e.g. due to a merger). With this second option the access points of the WAN and hotspots might provide some sort of additional feedback to users or to a separate networking entity (which we will call *Radio Access Controller* and abbreviate as RAC) operating across WAN and hotpsot networks. The task of the RAC would be to monitor the state of the network by collecting pertinent information from access points (such as, e.g. current utilization, channel rates available to users etc.) and dynamically coordinate the assignment of users' requests to various access points based on the collected network state information. As we already mentioned in Chapter 1, this architectural option is likely to provide a tighter control on the performance of the network, but is also likely to contribute to overall design complexity and cost, and might also lead to scalability problems.

Whichever architectural option is chosen, the way that the users are assigned to access points must affect the performance of the overall cooperative system. In this chapter we focus on understanding how the decision-making criteria and various parameters of heterogeneous network interplay and affect the performance. Our ultimate goal is to show how the controllable parameters in the system could be tuned so as to ensure the target performance under the least resource cost. Some important tunable parameters are the decision-metrics as well as resource allocation strategies e.g. network dimensioning.

It is important to understand that the design of decision-making metrics is a lot different for the two architectural options presented above. The decision-making entity in the "tightly coupled" heterogeneous network could be located within users' agents or within the RAC. The corresponding decision metric might involve a carefully designed feedback, (e.g. congestion levels of all APs in the network, amount of interference, proximity of end nodes to APs, etc) which the access points provide either to the end-nodes or the RAC. The decision-making design problem will then aim to find and tune the exact form of this feedback as well as the decision metric based on this feedback. In contrast, under the "loosely coupled" design option the decision-making has to undergo within the end nodes. Since the extra signaling from APs will be unavailable, the end-nodes would only be able to use some inferred information on the state of the network in their decision-making. This could be done, however, by observing the actual performance, e.g. typical service delays (in case of file transfers) and thus inferring the congestion levels at APs that the agent is able to communicate with. The design problem in this setting is hence finding good decision-making metric which utilizes no extra feedback from the APs.

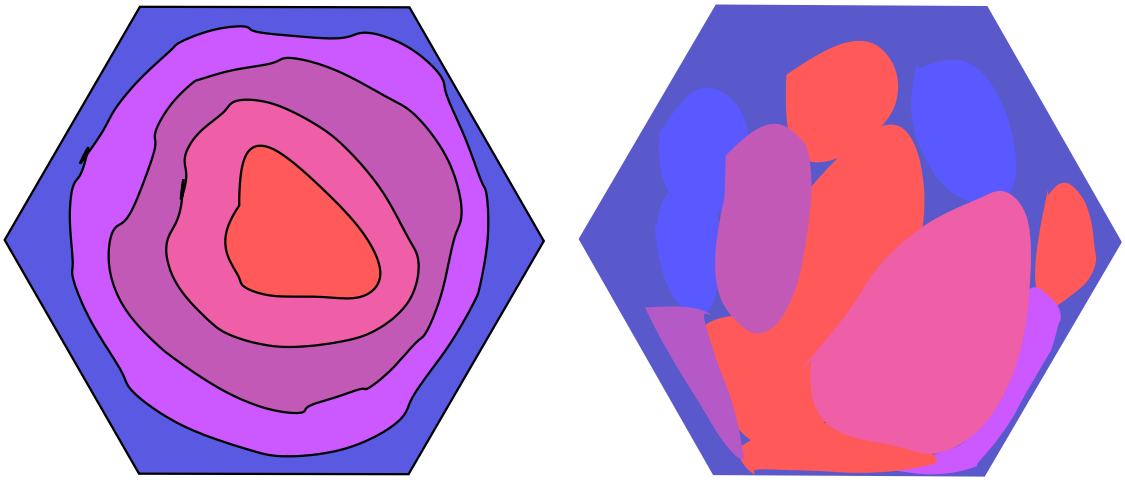
In this chapter we consider the design of heterogeneous networks under the "loosely coupled" option, leaving the "tightly coupled" case for consideration in Chapter 5. Note that under the "loosely coupled" option the users (or agents on the users' behalf) are in charge of

making selections between possible access points and thus we can use the network model that was analyzed in Chapters 2,3, where the decisions of individual agents are based on utility functions. Here we will assume that utilities are *exogenously* specified within a framework of a particular decision-making protocol, which we will design to optimize a particular performance objective.

We start by showing that, under some conditions, congestion-sensitive decision-making strategies are performance optimal. In particular, in Section 4.3 we show that if performance metric for each agent is given by a solely congestion dependent utility function, the equilibrium system performance will possess a max-min fairness property. In simple forms this property implies that in order to optimize the performance of the "worst" user per WAN service zone it is optimal to let agents make selections among the access points which lead to improvements in their individual performance. For example, if we need to optimize the network for data applications, it is optimal (under the conditions specified in the sequel) to let agents connect to access points which grant the least individual file transfer delays, since in that case the delays seen by the "worst" users in the WAN service zone will be the smallest. We verify these conclusions by performing several simulation experiments which illustrate the performance gains of congestion-sensitive decision-making strategies over more natural proximity-based strategies.

Good performance under congestion-dependent decision-making strategies is associated with an efficient usage of available network resources. Thus, for example hotspots are especially useful in spatial areas with a high and steady load, but can be under- or over- utilized whenever the loads in given spatial locations are time-varying and bursty (due to, e.g. daily user migration). At the same time the fluctuations at hotspots spatial scales can be "smoothed out" by the WAN network which is only sensitive to an aggregate load within much larger coverage areas. This smoothing property is realized by implementing congestion-sensitive decisions on agents, which will tend to prefer WAN service when they are within the coverage area of overly congested (relatively to the WAN) hotspot.

Once we demonstrate optimality of congestion-sensitive decision-making strategies, we focus on the complementary problem of minimizing available network resources while meeting a constraint on network performance. We solve a number of optimization problems that show how to optimally capacitate the network, and demonstrate that optimal capacity allocation might potentially significantly reduce the overall backhaul costs – backhaul links from WLAN access points to the wired network represent a significant fraction of the cost of operating such infrastructure [63]. Finally, we conclude Chapter 4 with Section 4.5 where we briefly identify the major takeaways from our analysis.



(a) **(b)**

Figure 4.1: **(a)**: Available WAN rates spatial profile. **(b)**: Traffic demands spatial profile.

4.2 Optimal heterogeneous network design problem

4.2.1 General intuition

Figure 4.1 (a) exhibits the possible spatial profiles for the realization of downlink bit-rates available from the WAN at different spatial locations within its service zone. Red color on the figure corresponds to higher available rates, while blue – to lower. Similarly, Figure 4.1 (b) shows a possible realization of average daily loads, with higher loads shown in red, and lower in blue. Clearly such information may and, in fact, as we will see in this chapter, should be accounted for when designing heterogeneous network.

In what follows we assume that the information on bit-rate and statistics of spatial loads is available for each WAN service zone – it could be obtained through measurements or could be deduced in some other way (e.g. predicted/projected). We will deal with several design problems. The first class of design problems assumes that the locations of hotspots are inflexible – they might have already been installed by a hotspot provider. For this class of problems we will still be able to optimize the backhaul bandwidth of hotspots – ideally the hotspots located in areas which are poorly covered by the WAN, or most densely populated will be provided with larger backhauls to sink more traffic. The second type of design problems assumes that the locations and the total number of hotspots (as well as the provided for them backhaul bandwidth) could be appropriately chosen.

In this chapter (as well as in the rest of the thesis) we will concentrate on wireless data services, with essential examples being FTP file downloads and web-browsing. Average delays seen by users during a typical file transfer are thus the natural quality of service metric in this setting, thus in this chapter we will consider two quality of service metrics that are related to the average delays which users experience.

For both types of design problems our objective will be to minimize the total backhaul bandwidth used within the WAN service zone under the constraint of providing a given amount of quality of service. Our motivation for considering backhaul dimensioning problems is that backhaul represents one of the largest recurring costs for operation of hotspots networks (see, e.g. [63]). For example, although bit-rates of up to 11 Mbps are available at current IEEE 802.11b systems, the practical achieved rates are rarely above 1-3 Mbps, since traffic is bottleneckhed by slower DSL/Cable providers' service rates.

Note that the selection of a decision-making strategy and backhaul dimensioning problems are tightly related to each other. Indeed, in one extreme scenario, agents might be configured to prefer the WAN at all times (because of, e.g. high-capacity WAN service at that area) making hotspots useless. In the opposite extreme, the agents could connect to hotspots exclusively when they are within hotspots coverage areaa – in that case the hotspots have to be equipped with sufficient backhaul bandwidth to ensure a reasonable quality of service. As we will show later such decision-making based on proximity may require large extra over-provisioning costs. This illustrates that optimization problem for backhaul dimensioning has to be considered in the context of a decision-making strategy.

4.2.2 Formalization of the problem

Consider the service zone S_m^w of a given WAN AP w_m. We will assume that H_m *non-overlapping, equal-sized* subzones (sites) are specified within S_m^w. We will denote these sites (in the same way as in Chapter 2) via $\{S_k^h\}_{k\in\mathcal{K}_m}$, and assume that S_k^h could be completely covered by the service zone of a single hotspot. These sites could represent the locations at which the hotspot provider has already installed hotspots, or the locations at which the hotspots are planned. A possible scenario is when $\cup_{k=1}^{H_m} S_k^h = S_m^w$, i.e. the (planned) hotspot sites cover the whole service zone S_m^w.

We will assume that the agents $a_i \in S_k^h$, share the same average (over time) WAN down-link data rate given by $B^w(a_i) = B^w(h_k)$, for $k \in \mathcal{K}_m$ and where $B^w(h_k)$ belongs to a discrete set $\mathcal{B} \triangleq \{b_r^w\}_{k=1}^{N_B}$ which consists of N_B distinct possible rates. We will not be specifying the up-link rates since our goal here is to consider heterogeneous network design for data services, which are mostly sensitive to the performance of the downlink. The major assumption that we take in this chapter, (but will dispense with in Chapter 5), is that the average over time down-link rates at each spatial location only depend on the large scale path loss, shadowing and short-term fading statistics, but are mostly insensitive to interference, e.g. out-of-cell interference from adjacent WAN APs. This will allow us to treat different WAN service zones as non-interacting, and thus will enable to optimize each the network on a per WAN cell basis. Our other major assumption is that all users within a hotspot's service zone are able to communicate with the hotspot's AP at a rate constrained by the bandwidth of the backhaul – as we already mentioned the wireless access bandwidth is likely to exceed by far that of the backhaul.

We will assume that the distribution of the number of agents within a site S_k^h is Poisson with density (per unit area) λ_k^a, where λ_k^a is possibly random itself. The number of agents at different sites S_k^h and S_l^h is assumed to be conditionally independent, given λ_k^a and λ_l^a, but the means λ_k^a and λ_l^a are possibly correlated. This construction will allow us to model scenarios with daily migration of users – in that case a number of contingent sites (e.g. covering a hotel lobby) will see loads varying around the same mean which slowly changes during the day.

Let us further specify the load and service models that we will be using. We will assume that each user generates a Poisson stream of download requests with intensity γ, each request is for files of average size f, and that all APs in the system serve users in a *processor sharing* fashion. This assumption works well [56] when modeling, for example, TDMA-based wireless systems in which all users connected to an AP receive an equal fraction of available time slots for service. A well-known current example of such systems is Qualcomm's 1XEV-DO (HDR).

We will denote by β_m a particular backhaul allocation strategy across sites S_k^h, for $k \in \mathcal{K}_m$ and WAN AP w_m and let $\beta \equiv \{\beta_m\}_{m=1}^M$. Let $\tau_m(a_i)$ denote the decision-making strategy that is used to assign a download request from an agent $a_i \in S_m^w \cap S_k^h$ to WAN AP w_m or hotspot AP h_k. We will also denote $\tau_m = \{\tau_m(a_i)\}_{a_i \in S_m^w}$ and $\tau = \{\tau_m\}_{m=1}^M$. Note that $\tau_m(a_i)$ in general can depend on many factors, such as proximity of agents to various APs, observed quality of channels, delays, etc. We will require $\tau_m(a_i)$ to be dependent only on information that is available locally to a_i, thus the decisions of an agent $a_i \in S_k^h$ can not, for example, explicitly depend on the current utilization of hotspot h_l, for $l \neq k$.

Let $P(\beta, \tau)$ denote a given performance metric in the system – this could be, for example, average delay in the WAN system or a worst user's average over time delay, etc. For conventional simplicity we assume that better performance corresponds to the case when metric P is smaller – this is, for example, the case when the metric is given by the mean delay seen by a typical user in the system. We formulate the general multi-provider network design problem as:

Problem 4.1. *(**Jointly optimal decision-making and backhaul dimensioning**):*

$$\min_{\beta,\tau} \left[\sum_{m=1}^{M} B_m^w + \sum_{k=1}^{K} B_k^h \right], \tag{4.1}$$

under constraints:

$$\mathbb{P}(P(\beta, \tau) \leq \theta) \geq 1 - \delta \tag{4.2}$$

$$0 \leq B_m^w \leq \hat{B}^w, \; B_k^h \geq 0. \tag{4.3}$$

In Problem 4.1 (4.2) ensures that performance in the system is at least a target θ with probability being only a small amount δ away from 1. The constraints (4.3) assume that the wireless access bandwidth at hotspots is unlimited and thus only constrained by the backhaul.

By contrast, only $\hat{B}^w$ is available at each WAN AP – usually WAN services operate using a limited portion of expensive licensed spectrum, and thus the wireless access bandwidth is likely to be the bottleneck.

4.2.3 Choice of performance metric

The explicit solution to Problem 4.1 is only possible when $P(\tau, \beta)$ is specified. As the choice of performance metric must depend on the application at hand and we focus on down-link data services, it seems natural for our purposes to relate performance to the delays that users experience during a typical file download. There are a lot of possibilities, however, in which performance could be related to the delay. For example the possible choices of performance metrics could be:

$$P(\beta, \tau) = \text{“delay, averaged over time and across users in the system”} \tag{4.4}$$

or

$$P(\beta, \tau) = \max_{m=1,\ldots,M} \Big(\text{“delay, averaged over time and across users within } S_m^w \text{”} \Big) \tag{4.5}$$

or

$$P(\beta, \tau) = \max_{a_i} \Big(\text{“delay, averaged over time for user } a_i \text{”} \Big) \tag{4.6}$$

and etc. Note that optimizing the system for each performance metric has its own advantages and disadvantages. Designs under metric (4.4) would ensure good performance on average but will tend to neglect some small fraction of users who experience relatively bad performance and which do not contribute much to the average delay in the system. A typical example of such “neglected” user would be the user not covered by any hotspot and falling on the boundary of WAN service zone, who thus sees relatively poor channel to WAN AP. To rectify this situation we could set our performance metric to (4.6) which favors users experiencing worst performance. In that case, however, under some pathological cases too much of network resources could be spent on providing service to a single user, while the rest of the network could benefit from these resources tremendously. We can illustrate this by getting back to the previous example with the user who is not covered by any hotspot and which experiences a very poor channel to the WAN (on average in time). Improving performance of such user would mean that the WAN (under no power control assumption) has to serve this user on average much longer than others. Since the WAN AP serves all users in processor sharing fashion, this situation can only realize if the decision-making of users covered by hotspots is biased in favor of hotspots, so that more of the WAN resources could be allocated to the user at the boundary of the WAN service zone. Overall this intuitively looks like a poor design choice, since most of network resources would be spent on a single user while they could be utilized more efficiently by the rest of the

users in the network.

This discussion illustrates that the choice of the performance metric should also be matched to the scenario conditions and be reasonably motivated. In what follows we will consider scenarios for which the WAN service is reasonably uniform, i.e. the average bit-rates experienced by agents at different spatial locations are not drastically different. For such scenarios we have selected the "worst user's performance" metric (4.6) to be the main measure of performance – partially due to the fact that with this choice of performance metric we are able to obtain closed-form results, and partially due to the fact that when the WAN service is "uniform enough" the design pathologies described in the above paragraph should not realize.

Let $D(a_i, \beta_m, \tau_m)$ denote the average over time delay that an agent $a_i \in S_m^w$ experiences when the decision-making strategies for agents within S_m^w is given by τ_m and backhaul within S_m^w is allocated according to strategy τ_m. In what follows for notational convenience we will often use $D(a_i)$ to denote the same quantity, keeping in mind that $D(a_i)$ depends on the strategies τ_m and β_m. In that case, using performance metric (4.6) Problem 4.1 reduces to solving the following optimization

Problem 4.2. *(***Design to achieve max-min fair performance***)*

$$\min_{\beta_m, \tau_m} \left[B_m^w + \sum_{k \in \mathcal{K}_m} B_k^h \right],$$

under constraints:

$$\mathbb{P}\left(\max_{a_i \in S_m^w} D(a_i, \beta_m, \tau_m) \leq \theta \right) \geq 1 - \delta$$

$$0 \leq B_m^w \leq \hat{B}^w, \; B_k^h \geq 0.$$

4.3 Optimal design when WAN coverage is uniform

We start by analyzing the case in which average bit-rates available at all spatial locations within the WAN service zone are the same – this is an over-simplified scenario which will gain us basic understanding of the main properties of solution to Problem 4.2. In reality, this scenario would correspond to the case when the WAN APs use large enough power to provide at least a target SINR at all spatial locations within their service zones, and the WAN APs do not adapt the modulation schemes (code rates) to operate on the peak available channel capacity.

Let B_m^w denote the (common) WAN link-rate seen by all agents within S_m^w. In this scenario, let us also assume that all hotspots within S_m^w use same bandwidth at the backhaul. By using a well known result from queueing theory (see, e.g. [57]) we conclude that the average

delay seen by an agent a_i connected to the WAN is given by:

$$D(a_i) = \frac{1}{f/B_m^w - \gamma N_m^w}, \tag{4.7}$$

where, as in Chapter 2 N_m^w denotes the total number of users connected to WAN AP w_m. Similarly, the average delay seen by an agent $a_j \in S_k^h$ connected to hotspot's AP h_k is given by:

$$D(a_j) = \frac{1}{f/B_k^h - \gamma N_k^h}. \tag{4.8}$$

Note that in equations (4.7,4.8) we have neglected any boost in performance which might have resulted from APs serving users in opportunistic fashion [56] – this step will greatly simplify our analysis. One can also argue that the effect associated with opportunistic scheduling could be modelled by introducing a (typically slow varying with the number of users) multiplicative factor which reflects any increase in the effective observed service rates. Thus, for example, if channel fading is Rayleigh, then this multiplicative factor is roughly proportional to the log of the total number of users connected to an access point, and therefore we can incorporate it by scaling B_m^w (or B_k^h) accordingly.

4.3.1 Optimal decision making

Note that one can make a utility of an agent to be negative of the average delay that an agent experiences when it connects to a WAN or hotspot's AP. This of course, assumes that each agent is capable to estimate it with sufficient precision – here we will always assume that this is indeed the case. According to Definition 2.2 and under the "uniform WAN service" approximation such utilities have a solely congestion structure – at least for moderate loads[1]. One can also verify that the equilibria corresponding to such utilities are always unique – thus we could use the properties of unique equilibria identified in Section 2.6. In particular we can state the following corollary to Proposition 2.7:

Corrollary 4.1. *For the case of "uniform WAN service", decision-making based on greedy minimization of average delay seen by an agent is optimal for Problem 4.2.*

Corollary 4.1 implies that backhaul dimensioning in Problem 4.1 must be done under the assumption that agents connect to access points in an effort to reduce their experienced average delay. Here, to confirm the conclusion of Corollary 4.1 we will report the results of several simulations, where we compare the performance of a heterogeneous network operating under

[1]The utilities based on average delay are not defined for all $N_m^w, N_k^h > 0$ – for excessive loads delays become infinite. However this poses no problem for decision-making model and subsequent analysis, since no users will tend to be choosing an AP where the delays become significantly large.

Table 4.1: Simulation parameters

Parameter	Notation	Value
Agents' density	λ^a	100 km^{-2}
Hotspots' density	λ^h	30 km^{-2}
WAN APs' density	λ^w	0.81 km^{-2}
Hotspot coverage radius	d	100 m
Requests/user/s	γ	1 min^{-1}
Average file size	f	80 kB
Simulated area	$\|A\|$	100 km^2
WAN b/w	$B^w(a_i)$	1 Mbps

congestion-sensitive decision-making based on negative delays in place of utilities with the performance of such system under proximity-based decision making in which all agents connect to hotspots when covered by the corresponding hotstpots' service zones.

Illustrating simulation example. The simulation example that we discuss here uses the geometric model constructed in Section 2.3.2. The locations of agents, hotspots and WAN APs are assumed to be given by independent Poisson point processes with densities λ^a, λ^h and λ^w, the values of which are specified in Table 4.1 along with the values for other system parameters. We consider two performance metrics. The first is the mean delay averaged across users – we refer to it as "average performance" and it is defined as:

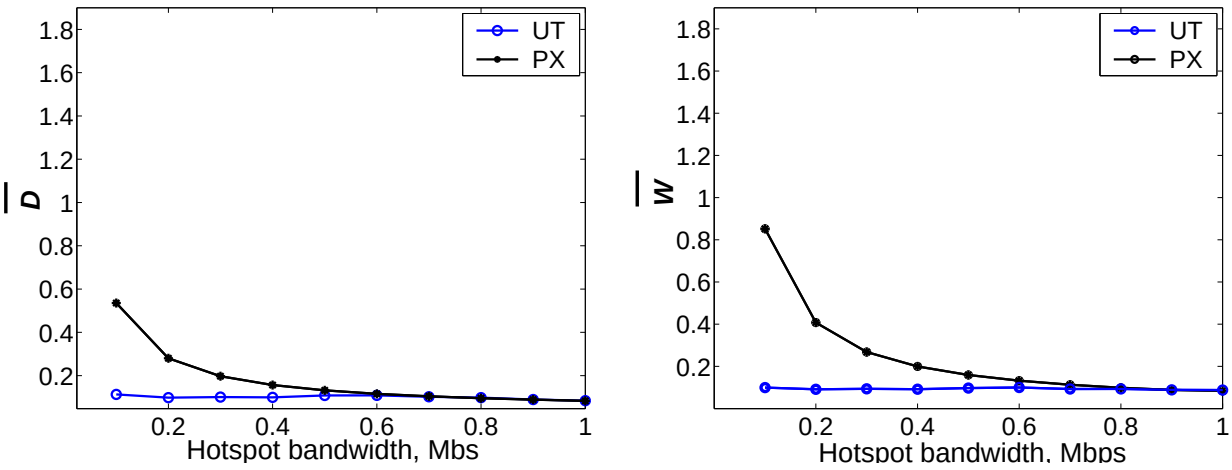
$$\bar{D} \triangleq \frac{1}{|\tau^a(A)|} \sum_{a_i \in \tau^a(A)} D(a_i)\,,$$

where A is the simulated area and $D(a_i)$ is average file transfer delay seen by agent a_i. The second metric is the average worst case user's delay per WAN service zone, and is defined as:

$$\bar{W} = \frac{1}{|\tau^w(A)|} \sum_{w_m \in \tau^w(A)} \max_{a_i \in S_m^w} D(a_i)\,,$$

for a simulated region A. Since we do not have blocking, all our results are conditioned on the event that the overall system is stable. However, the simulation parameters are chosen in such a way that the probability of instability is very small.

Figure 4.2 shows $\bar{D}$ and $\bar{W}$ after convergence to equilibrium versus the hotspot bandwidth B^h for utility based (UT) and proximity based (PX) selection strategies. There are significant gains both in the average per user and worst case performance per cell if the available bandwidth at hotspots is less than 60% of that available at the WAN. Given the parameters in Table 4.1 one might deduce that PX based strategy needs at least five times more bandwidth at



Average performance

Average worst per cell performance

Figure 4.2: Performance gains of congestion-sensitive decision-making based on delay (UT) over simple proximity-based decision-making (PX) in a scenario with uniform WAN coverage.

the backhaul than the UT strategy to achieve the same average per user performance and even more bandwidth to achieve the same value of $\bar{W}$.

The results of this experiment confirm the conclusion of Corollary 4.1. The performance gains arise from the ability of WAN APs to cover considerably larger service areas than hotspots and thus statistically multiplex spatial load fluctuations across a large number of much smaller hotspots coverage areas. Thus the WAN APs can serve as a pooled resource which absorbs load fluctuations within hotspots, and such pooling could be realized by imposing congestion-sensitive decisions on users' agents.

4.3.2 Backhaul dimensioning

Next we proceed to solving the remaining part of Problem 4.2 which considers the optimal allocation of backhaul bandwidth across hotspots. We consider first the case in which the locations of hotspots have already been specified – these might correspond to the existing locations for hotspots' business, e.g. the locations of coffee-shops. In that case we can explicitly solve for optimal backhaul that is required at hotspots and WAN APs – the solution is presented in Proposition 4.1 which follows below.

Let $\hat{M}_m^w(\delta)$ be the largest integer such that:

$$\mathbb{P}(M_m^w > \hat{M}_m^w(\delta)) \leq \delta,$$

and

$$\hat{N}_m^w(\theta) = \left\lfloor \frac{\hat{B}^w}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor. \tag{4.9}$$

Here $\hat{N}_m^w(\theta)$ is the the largest number of agents that the WAN AP could serve with delay not exceeding θ, if $\hat{B}^w$ was allocated to this AP. We have the following proposition:

Proposition 4.1. *Let the user spatial densities λ_k^a be identically (but, possibly, not independently) distributed across $k \in \mathcal{K}_m$. Then, there are three regimes to consider in solving Problem 4.1: (i) If*

$$\hat{N}_m^w(\theta) \geq \hat{M}_m^w(\delta), \tag{4.10}$$

then a policy that allocates $B_m^w = \gamma f \hat{M}_m^w(\delta) + \frac{f}{\theta}$ to the WAN AP w_m and no bandwidth to any of the hotspots is optimal for Problem 4.1 when H_m is large enough.
(ii) If

$$\mathbb{P}(M_{\bar{C}_m}^w > \hat{N}_m^w(\theta)) > \delta, \tag{4.11}$$

then no solution to Problem 4.1 exists.
(iii) If both (4.10) and (4.11) are violated, then a policy that is optimal for large enough H_m allocates $B_m^w = f\gamma \hat{N}_m^w(\theta) + f\theta$ units of bandwidth to the WAN AP and $B^h(\theta, K) = f\gamma K + f\theta$ units of bandwidth to each of the hotspots in S_m^w, where K is the smallest integer such that:

$$\mathbb{P}\left(M_{\bar{C}_m}^w + \sum_{k \in \mathcal{K}_m} (M_k^h - K)\mathbf{1}_{\{M_k^h > K\}} > \hat{N}_m^w(\theta) \right) \leq \delta. \tag{4.12}$$

Proof. Observe, that marginal distribution of the number of agents in a hotspot h_k is the same for all $k \in \mathcal{K}_m$, thus by symmetry we must allocate the same amount of bandwidth to each of the hotspots. To show the optimal allocation for the first regime, consider an allocation strategy $\mathcal{A}_1$, where each of the hotspots is given B^h units of bandwidth. Then a hotspot could serve at most $N^h(\theta, B^h)$ agents, where

$$N^h(\theta, B^h) = \left\lfloor \frac{B^h}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor.$$

The total number of agents which hotspots could serve is:

$$N_1 = H_m \left\lfloor \frac{B^h}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor \leq \frac{H_m B^h}{\gamma f} - \frac{H_m}{\theta\gamma}. \tag{4.13}$$

Now consider an allocation strategy $\mathcal{A}_2$ that shifts ΔB units of bandwidth from each hotspot to the WAN AP, where $\Delta B < B^h$ is such that

$$\frac{B^h - \Delta B}{\gamma f} - \frac{1}{\theta\gamma} \in \mathbb{N}.$$

We will assume that the WAN AP uses the shifted bandwidth to serve the agents within the hotspots, Then, from (4.1) and (4.2) the total number of agents in C_m that could be served by

such system, without violating the delay requirement is:

$$N_2 = H_m \left\lfloor \frac{B^h - \Delta B}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor + \left\lfloor \frac{H_m \Delta B}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor \geq H_m \frac{B^h - \Delta B}{\gamma f} - \frac{H_m}{\theta\gamma} + \frac{H_m \Delta B}{\gamma f} - \frac{1}{\theta\gamma} - 1$$
$$= \frac{H_m B^h}{\gamma f} - \frac{H_m + 1}{\theta\gamma} - 1 \geq N_1 - 1 - \frac{1}{\theta\gamma}. \tag{4.14}$$

Now note that under allocation strategy $\mathcal{A}_1$, N_1 agents in total could be served in the hotspots only if each of them in fact had $N^h(\theta, B^h)$ agents to serve. However, if the system gets large enough ($H_m \gg 1$) with probability arbitrary close to 1 there is at least $\left\lceil 1 + \frac{1}{\theta\gamma} \right\rceil$ hotspots containing at most $N^h(\theta, B^h) - 1$ agents at their service zones. We thus obtain that the performance under allocation strategy $\mathcal{A}_2$ is at least as good as the performance under $\mathcal{A}_1$ once H_m is large enough. This shows that if (4.10) holds then a policy that allocates $B_m^w = \gamma f \hat{M}_m^w(\delta) + \frac{f}{\theta}$ to the WAN AP w_m is optimal for sufficiently large H_m.

Consider the regime where the inequality (4.10) is not satisfied. Then, a policy which is optimal for large enough H_m allocates $B_m^w = f\gamma \hat{N}_m^w(\theta) + f\theta$ units of bandwidth to the WAN AP, where $\hat{N}_m^w(\theta)$ is given by (4.9). For a particular realization of agents let the number of agents that do not fall within a service zone of any of the hotspots be denoted as $M_{\bar{C}_m}^w$ and let

$$\Delta N_m^w = \hat{N}_m^w(\theta) - M_{\bar{C}_m}^w. \tag{4.15}$$

If $\Delta N_m^w > 0$ then ΔN_m^w agents can be served by WAN AP in any of the hotspots without violating the delay constraint at the WAN AP. Clearly, the optimal way to use the extra bandwidth is to serve agents from the most congested hotspots. In particular, assuming that agents in C_m are initially connected to their hotspots, an algorithm that selects which agents the WAN would serve, at each step takes the most congested hotspot and switches an agent connected to this hotspot to the WAN AP. As a result, the number of agents connected to hotspots and the WAN will be represented by the Figure 2.3 with a "slicing" plane at some level.

If, however, for some realization of agents $\Delta N_m^w < 0$ then no agents inside the hotspots' service zones can be served by the WAN. In this case, the agents in $\bar{C}_m$ connected to the WAN do not meet their delay requirement. Denote by F^w the event:

$$F^w(\theta) = \{\hat{N}_m^w(\theta) < M_{\bar{C}_m}^w\}. \tag{4.16}$$

Clearly if $\mathbb{P}(F^w) > \delta$, then the optimization problem (4.1) does not have a solution that meets the probabilistic requirement (4.2) (statement (ii) of the proposition).

Thus, in the remaining case, we assume that $\mathbb{P}\big(F^w(\theta)\big) < \delta$, and hence the delay requirement of the agents in $\bar{C}_m$ is always met. Below we find the minimum bandwidth that has to be allocated to hotspots so that the agents within C_m meet their delay requirement too. We fix

$K > 0$ and let:

$$B^h(\theta, K) = f\gamma K + f\theta\,.$$

Thus $B^h(\theta, K)$ is the amount of bandwidth that each hotspot has to be supplied to serve up to K agents within its service area. Assume that $B^h(\theta, K)$ is indeed provided to each of the hotspots, then the event that any hotspot h_k has more then K agents connected to it is equivalent to the event $F^h(\theta)$ which has:

$$\sum_{k \in \mathcal{K}_m} (M_k^h - K)\mathbf{1}_{\left\{M_k^h > K\right\}} > \Delta N_m^w\,, \tag{4.17}$$

Clearly the event $F^w(\theta)$ implies the event F^h, and thus, guaranteeing that $F^h(\theta)$ does not occur is enough to guarantee that F^w does not occur either. Thus, via plugging (4.15) into (4.17) we have that the value of K is as given in part (iii) of the proposition. □

Note that Regime (iii) identified by Proposition 4.1 can be viewed as the regime where both WAN and hotspots benefit from cooperation. Indeed, the WAN AP is unable to handle all the traffic due to the limit on the wireless access bandwidth. At the same time the backhaul allocated with each hotspot enables a hotspot to serve at most K agents, where K is given by (4.12). It is optimal to shift the "overload" in each hotspot to the WAN AP. As the Corollary 4.1 shows, in order to realize such load shifting in a distributed fashion, one needs only to implement a greedy average-delay based decision-making for the agents.

Backhaul cost savings under optimal dimensioning. In what follows we compare the optimal total bandwidth in the sense of Proposition 4.1 with the total bandwidth that would be required to meet the probabilistic constraint (4.2) if the system was designed under a proximity-based decision-making mechanism. To simplify exposition, we will make this comparison under the assumption of having no upper constraint on the bandwidth that is used by the WAN AP, i.e. let $\hat{B}^w = \infty$.

As seen earlier, the optimal strategy for Problem 4.1 allocates sufficient resources on the WAN and allows all agents to connect to WAN APs. Recall that under PX strategy the agents falling within the service zones of the hotspots must connect to the hotspots. Under both optimal and PX resource allocation the agents that do not fall within the service zones of any hotspot must be served by the WAN, thus there is a comparable cost for both PX and optimal strategy that is associated with provisioning at the WAN for this type of agents. At the same time, we expect to see overprovisioning costs associated with agents at hotspots to be quite large for PX strategy in comparison to the optimal.

We find a lower bound on the savings in overprovisionning by considering a suboptimal strategy that allocates two separate channels for agents that are within and outside C_m. Under this strategy, bandwidth cost associated with meeting a delay requirement for the users in $\bar{C}_m$

is exactly the same as for PX strategy. We are left to compare only the savings in bandwidth associated with serving the agents in C_m by either hotspots or the WAN AP.

We will find the minimal bandwidths B_{PX} and B_o that is required to meet the delay requirement (4.2) of agents within C_m, once the PX or optimal connection strategy respectively is deployed. We define the access bandwidths:

$$\Delta B_{PX} = B_{PX} - \bar{B}\,, \text{ and } \Delta B_o = B_o - \bar{B}\,.$$

Here $\bar{B}$ is the minimal bandwidth that has to be used to serve the agents in C_m when there are exactly average number of them, $H_m\lambda^a\tau d^2$, residing in C_m, thus

$$\bar{B} = \gamma f H_m \lambda^a \tau d^2 + \frac{f}{\theta}\,.$$

Proposition 4.2. *Let the user densities λ_k^a for $k \in \mathcal{K}_m$ be deterministic and equal for all $k \in \mathcal{K}_m$. Then, for large H_m, the overprovisionning costs scale as:*

$$\frac{\Delta B_o}{\Delta B_{PX}} = O\left(\frac{1}{\sqrt{H_m}}\right).$$

Proof. We first find the minimum bandwidth that has to be allocated to hotspots to meet the delay requirement (4.2). Denote $\bar{M} = \mathbb{E}\left[M_k^h\right]$ and let $\sigma^2 = \mathbf{var}[M_k^h] = \mathbb{E}\left[M_k^h\right]$. Let $\kappa^h(\delta, H_m)$ be the smallest positive number such that:

$$\left[\mathbb{P}(M_k^h \le \bar{M} + \kappa^h(\delta, H_m)\sigma)\right]^{H_m} \ge 1 - \delta\,. \tag{4.18}$$

Clearly, $\kappa^h(\delta, H_m)$ is a nondecreasing function of H_m for any fixed δ. From (2.4), we obtain:

$$\mathbb{P}\left(\max_{a_i \in S_m^w} D(a_i) > \theta\right) \le \delta\,,$$

if and only if

$$B_{PX} = B_{PX}(\delta, H_m, \theta) = \gamma f(\bar{M} + \kappa^h(\delta, H_m)\sigma) + \frac{f}{\theta}\,.$$

Thus $H_m B_{PX}(\delta, H_m, \theta)$ is the minimum total bandwidth that has to be allocated for hotspots when PX strategy is deployed, which gives the excess bandwidth:

$$\Delta B_{PX} = (H_m - 1)\frac{f}{\theta} + H_m \kappa^h(\delta, H_m)\sigma\,.$$

Now we find the total bandwidth that would be needed by the WAN AP to serve the agents within the hotspots and meet the delay requirement. Following the same logic as above,

we find:

$$B_0 = B_0(\delta, H_m, \theta) = \gamma f(H_m \bar{M} + \kappa^w(\delta)\sigma\sqrt{H_m}) + \frac{f}{\theta},$$

where we defined $\kappa^w(\delta)$ as:

$$\mathbb{P}(\sum_{k \in \mathcal{K}_m} N_k^h \leq H_m \bar{M} + \kappa^w(\delta)\sigma\sqrt{H_m}) = 1 - \delta.$$

Note that, for sufficiently large H_m we can assume that the Central Limit Theorem holds and then the sum $\sum_{k \in \mathcal{K}_m} M_k^h$ is distributed normally with variance $H_m\sigma$. Since the mean and the variance uniquely define any normal distribution, we have that $\kappa^w(\delta)$ does not depend on H_m. Therefore, for sufficiently large[2] H_m and any fixed $\delta > 0$ we have $\kappa^w(\delta) \leq \kappa^h(\delta, H_m)$. The excess bandwidth when all agents in C_m are served by the WAN is given by:

$$\Delta B_o = B_o - \bar{B} = \kappa^w(\delta)\sigma\sqrt{H_m}.$$

Comparing ΔB_m^w and ΔB^h we find:

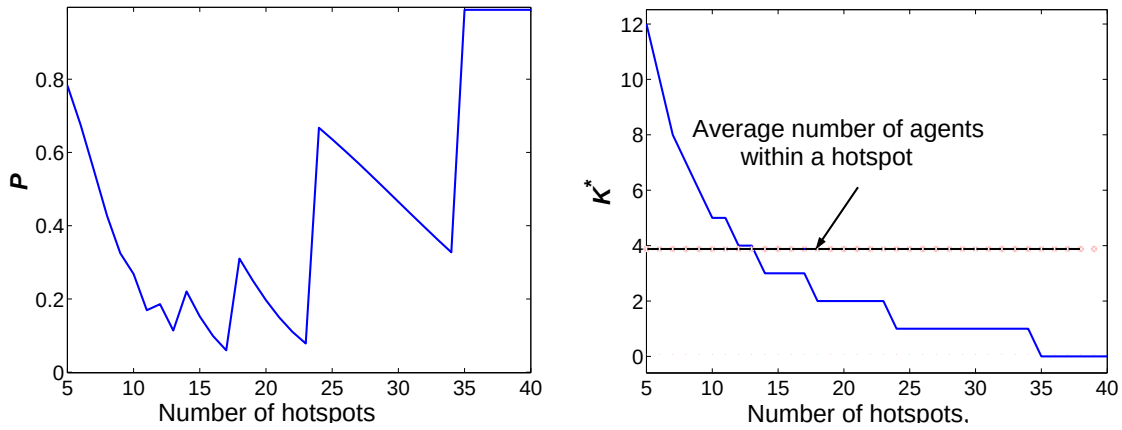
$$\frac{\Delta B_o}{\Delta B_{PX}} = O\left(\frac{1}{\sqrt{H_m}}\right), \tag{4.19}$$

where we used that $\kappa^h(\delta, H_m) \leq \kappa^w(\delta)$ for sufficiently large H_m. □

Remark 4.1. *Note that when the delay requirement is very stringent, in particular when* $\theta \ll (\gamma\lambda|S_k^h|)^{-1}$*, then the excess bandwidth* $\Delta B_{PX} \gg \bar{B}$*. Since* B_o *is of the same order as* $\bar{B}$ *we may find that the scaling of Proposition 4.2 holds also when the excess bandwidths* ΔB_o *and* ΔB_{PX} *are replaced by total bandwidths* B_o *and* B_{PX} *respectively.*

It is important to note that the exact cost of overprovisionning might vary depending on the realized scenario conditions. Proposition 4.2 estimated these costs for the "mild" case when the number of users in each hotspot is identically and *independently* distributed for $k \in \mathcal{K}_m$. More profound overprovisioning costs are observed in some extreme scenarios, in which the density of users is correlated across hotspots. Such scenarios are quite typical in reality – for example users might be concentrated in one part of the WAN zone, e.g. during work hours, but could migrate to other parts of the WAN zone during lunch breaks, etc. Thus, for example it is quite possible that users, originally uniformly distributed within a WAN service, get to the same hotspot during a lunch break. Clearly, to ensure good performance under proximity based decision strategy one would need a tremendous resources to be allocated to cover the traffic at the hotspot, and these resources would be completely wasted during the off-peak hours. In

[2] In fact we have checked numerically that for $1 \leq \lambda \leq 100$ and $0 < \delta < 1$ we have $\kappa^w(\delta) \leq \kappa^h(\delta, H_m)$ already when $H_m \geq 3$.



(a) **(b)**

Figure 4.3: **(a)**: The largest number of agents a hotspot could serve without violating target delay vs. H_m. **(b)**: Probability of exceeding target delay vs. H_m.

contrast, under congestion-sensitive decision-making, the hotspot would require some moderate amount of backhaul, whereas the overloads during the peak hour could be "shifted" towards the WAN.

Optimal number of hotspots for a given backhaul size. Proposition 4.1 could also be used to compute the optimal number of hotspots that should be placed within S_m^w. In particular a service provider might wonder if putting more hotspots in the area but having them able to support fewer users is better then doing otherwise. Indeed, we have a tradeoff between the risk associated with having users uncovered by a hotspot and the risk of often having users that hotspots can not support.

By Proposition 4.1, part (iii) we know that B_m^w should be made as large as the available spectrum can support, and thus we have to decide only on how many hotspots would give best performance within S_m^w when there is a constraint on the total amount of bandwidth used by hotspots backhaul, B_{tot}^h. More formally, we need to find H_m, such that

$$\mathbb{P}\left(M_{C_m}^w + \sum_{k\in\mathcal{K}_m}(M_k^h - K)\mathbf{1}_{\{M_k^h>K\}} > \hat{N}_m^w(\theta)\right) \tag{4.20}$$

is minimized under constraint:

$$H_m(f\gamma K + f\theta) \leq B_{tot}^h\,. \tag{4.21}$$

Note that behavior of the probability given by (4.20) as a function of H_m might be quite complex. Figure 4.3 (a) exhibits this behavior in a typical scenario. Observe that P in the graph 4.3 (a) goes down steadily with H_m, until $H_m = 10$. At this point, each hotspot can support just above average number of agents, falling within their service zones, without violating the

delay requirement (see Figure 4.3 (b)). The graph is much less regular once each hotspot is able to support less then the average number of agents that fall within their service zones. We note, that the lowest point at the graph is for $H_m = 17$, whence the number of agents that could be supported by each of the hotspots is below the average number of agents that fall within a hotspot.

4.4 Optimal design in the case of non-uniform WAN coverage

Let us now assume that the WAN AP does not provide uniform quality of service to all spatial locations within its service zone. This situation would occur if, for example, the signal from the WAN AP at a particular location is shadowed by an obstruction. In this scenario users at different spatial locations will perceive the WAN service differently, and are likely to benefit from hotspots in different ways. The objective of this section is to provide a recipe for decision-making design and complementary backhaul dimensioning, which are suboptimal in the sense of Problem 4.1, but still provide a reasonable design optimization starting point, and as a result will realize significant performance gains and reduce infrastructure costs.

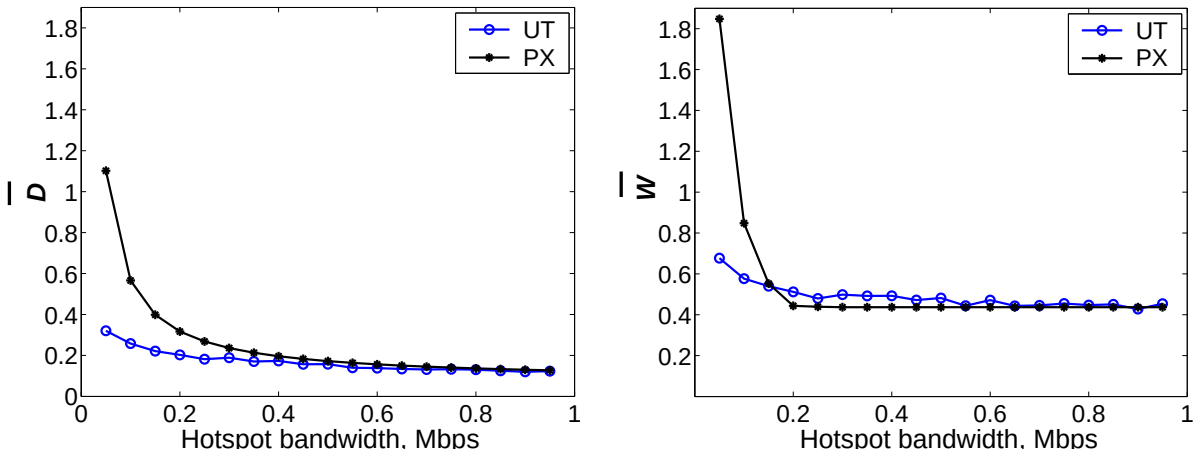
4.4.1 Performance under a congestion-sensitive strategy

Note that the optimality of decision-making based on greedy minimization of the individual users' delays ceases to hold in a scenario with a non-uniform WAN coverage. Firstly, the delays of agents connected to the WAN are now given as those in a multi-class $M/GI/1-PS$ queue, e.g. for an agent a_j connected to the WAN AP w_m at time t we have:

$$D^w(a_j) = -\frac{f}{B^w(a_j) - \gamma f \sum_{a_i \in \mathcal{W}_m(t)} B^w(a_j)/B^w(a_i)}, \tag{4.22}$$

where $\mathcal{W}_m(t)$ is the set of all agents connected to w_m. Thus, if we associate the utility of an agent a_j connected to the WAN with the average delay via $U_j^w \equiv -D^w(a_j)$, the utility U_j^w would not have a solely-congestion structure. Moreover, U_j^w does not belong, in general to the class of congestion and agent dependent utilities, since it depends not only on the total number of agents connected to w_m, but also on the WAN rates available at locations of all the agents connected to the WAN AP w_m.

It is still intuitively clear, that as long as the WAN service is "mostly" uniform, the congestion sensitive decision-making should perform fairly well. To confirm this intuition, on Figure 4.4 we plot the results of an experiment where we have used the same setup as in the example of Section 4.3.1 (see Table 4.1), except that the SINR and, correspondingly, the WAN rate seen at each spatial location is simulated as a random quantity. In particular, the SINR is



Average performance

Average worst per cell performance

Figure 4.4: Performance of congestion-sensitive decision-making based on delay (UT) and a proximity based decision-making, in a scenario with non-uniform WAN coverage.

assumed to be given via the SINR distribution shown at Figure 4.5 (taken from [56]) and the corresponding rate is mapped from SINR similarly to the way it is done in Qualcomm 1XEV-DO system [66] (see the table shown at Figure 4.5).

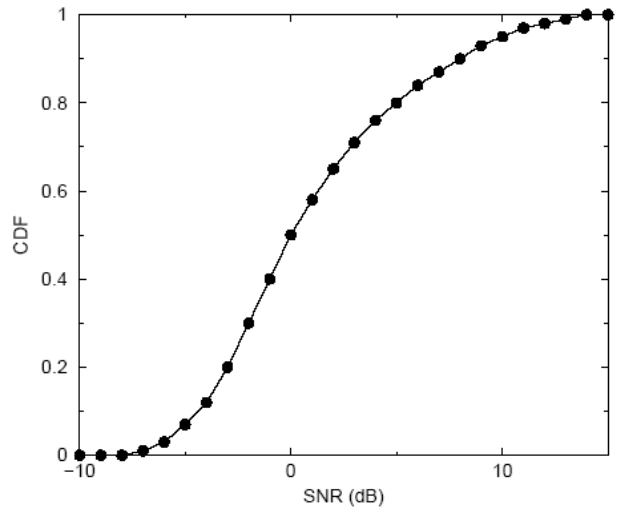
Note that the performance shown in Figure 4.4 is the performance after the decision-making dynamics has converged to an equilibrium. Since the utilities of agents connecting to the WAN are more complex than the ones for which we established such convergence, the fact that the dynamics indeed converge to an equilibrium is quite remarkable. However, whenever the size of the WAN service zone is large enough to contain a lot of agents, and the rates for all agents are identically distributed, we have:

$$\sum_{a_i \in \mathcal{W}_m} 1/B^w(a_i) \sim N_m^w \mathbb{E}\left[1/B^w(a_j)\right] + O\left(\sqrt{N_m^w}\right),$$

which indicates that the two most important factors affecting U_j^w are the total number of agents connected to the WAN AP, N_m^w and the average WAN downlink rate, $B^w(a_j)$, seen at the agent's spatial location. One can thus expect the system dynamics to be close to the one where the utilities U_j^w are of congestion and agent dependent type.

Observe that, under both performance metrics, Figure 4.4 indicates that the performance of congestion-sensitive decision strategies is superior to that of proximity based strategies, but only when the backhaul resources at hotspots are limited. The graph for $\bar{W}$ on Figure 4.4 could also be used to deduce that the decision-making based on greedy delay minimization does not, in general result in a max-min fair configuration of agents' equilibrium choices.

Motivated by the results of Section 4.3 and the intuition presented above, we will not attempt to solve for the optimal decision-making strategy (in the sense of Problem 4.1), but



SNR $\geq$	Rate (Kbs)
-12.5	38.4
-9.5	76.8
-8.5	102.6
-6.5	153.6
-5.7	204.8
-4.0	307.2
-1.0	614.4
1.3	921.6
3.0	1228.8
7.2	1843.2
9.5	2457.6

Figure 4.5: Qualcomm SNR cdf (left) and SNR-data rate correspondence (right).

will stick to the decision-making based on greedy delay minimization of each user. In Problem 4.1 we thus have to solve for the backhaul allocation among hotspots, which is likely to have a spatially dependent structure due to the difference in WAN rates seen in different spatial locations.

4.4.2 Backhaul dimensioning

Problem 4.2 is much more complex in a general scenario with spatially dependent WAN rates and spatially dependent user densities. We will make several simplifying assumptions and reformulate the problem so as to retain its key aspects. The following are our major assumptions:

1. We assume that the spectrum at the WAN is fully utilized. Since the WAN service could be arbitrarily poor in some locations it might happen that a single hotspot installed at such locations would exploit the backhaul bandwidth much more effectively. Thus the solution to an analog of Problem 4.1 posed in this scenario might lead in some cases to the conclusion that the available at the WAN spectrum for communication should not be fully utilized. Since the cost of purchased spectrum probably exceeds by far the backhaul associated costs within a cell, such a solution would indicate that the placement of the WAN AP or overall WAN design is poor.

2. We assume that the operation regime is such that each hotspot takes the largest number of agents within its service zone that it can serve with average delay of at most θ and "shifts" the remaining agents to the WAN AP. (Recall that θ is the target delay-performance threshold in Problem 4.2)

3. Our last modification to Problem 4.1 is that in place of $\max_{a_i \in S_m^w} D(a_i)$ we will concentrate on the delay averaged across agents that are connected to the WAN AP w_m:

$$\bar{D}_m^w = \frac{1}{N_m^w} \sum_{a_j \in \mathcal{W}_m} D(a_j)\,,$$

where $\mathcal{W}_m$ denotes the set of agents that are connected to the WAN via the rule described in 2. Using the expression for average delay in the multi-class M/GI/1-PS queue we have:

$$\bar{D}_m^w = \frac{1}{N_m^w} \left(\frac{1}{\sum_{a_j \in \mathcal{W}_m} \frac{f}{B^w(a_j)}} - \gamma \right)^{-1}. \tag{4.23}$$

Assumption 1 allows us to consider optimization of backhaul allocation associated only with hotspots, and Assumption 2 permits us to account only for the performance of agents connected to the WAN AP. With the set of Assumptions 1–3 we arrive at:

Problem 4.3.

$$\min_{\{B_k^h\},\, k \in \mathcal{K}_m} \left[\sum_{k \in \mathcal{K}_m} B_k^h \right], \tag{4.24}$$

under constraints:

$$\mathbb{P}\left(\bar{D}_m^w \leq \theta\right) \geq 1 - \delta, \tag{4.25}$$

$$B_k^h \geq 0\,,\ \forall k \in \mathcal{K}_m\,. \tag{4.26}$$

Note that the constraint (4.25) in Problem 4.3 is a relaxed version of the respective constraint (4.2) in Problem 4.1, since clearly, (4.2) implies (4.25) (but not vise versa). It is intuitively clear that the solution to Problem 4.3 will tend to neglect a small fraction of users that experience bad performance but do not contribute much to the average (per WAN service zone) delay.

Let us assume that each hotspot h_k, $k \in \mathcal{K}_m$ is provided enough bandwidth to serve up to K_k agents with delay not exceeding θ, i.e.:

$$B_k^h = \gamma f K_k + \frac{f}{\theta}\,.$$

Solving Problem 4.3 reduces to finding the optimal set of values $\{K_k\}_{k \in \mathcal{K}_m}$, where $K_k \geq 0$ for all $k \in \mathcal{K}_m$. Proposition 4.3 shows how to approximate the values for K_k in a simple case where the distribution of users within S_m^w is given by a homogeneous Poisson process. Note that without loss of generality we assumed $\cup_{k \in \mathcal{K}_m} S_k^h = S_m^w$, since for a site h_l which can not be used for hotspot installation we can always set $K_l = 0$.

In Proposition 4.3 which follows below, we approximate the distribution of the number of agents within S_k^h via a Gaussian random variable η such that $\mathbf{var}[\eta] = \mathbb{E}[\eta] = E[M_k^h]$. Then

$$g(K_k) \triangleq \mathbb{E}\left[(\eta - K_k)\mathbf{1}_{\{\eta > K_k\}}\right],$$

approximately gives the average number of agents within a hotspot that connect to the WAN and

$$L\left(\{K_k\}_{k\in\mathcal{K}_m}\right) \triangleq \sum_{k\in\mathcal{K}_m} g(K_k), \tag{4.27}$$

approximately gives the average of the total number of agents within S_m^w that connect to the WAN. We replace N_m^w in (4.23) via its average approximated by (4.27) and treat K_k for each $k \in \mathcal{K}_m$ as taking continuum values. This allows us to reduce Problem 4.3 to a nonlinear programming one, at which point we use Kuhn-Tucker conditions to arrive at the approximate solution for the set $\{K_k\}_{k\in\mathcal{K}_m}$.

Proposition 4.3. *Let the distribution of users within S_m^w be given by a (homogeneous) Poisson spatial process with constant density. We have that either $K_k = 0$, or*

$$\frac{\mathbb{P}(\eta > K_k)}{B_m^w(h_k)} = \nu,$$

for some constant ν, such that the set $\{K_k\}_{k\in\mathcal{K}_m}$ obeys:

$$\sum_{k\in\mathcal{K}_m} \frac{f}{B_m^w(h_k)} g(K_k) = \left(\frac{1}{\theta L\left(\{K_k\}_{k\in\mathcal{K}_m}\right)} + \gamma\right)^{-1}.$$

Remark 4.2. *Note that the solution for K_k does not include any dependence on δ. This, in fact is a result of us neglecting certain terms, which depend on δ, but are provably much smaller than the terms we kept, independent of δ. For more details, see the following proof.*

Proof. We will derive the approximate solution to Problem 4.3 under the assumption[3] that $K_l = K_k$ when $b_l^w = b_k^w$, and $l, k \in \mathcal{K}_m$. First, we will elaborate on the expression (4.23) for $\bar{D}_m^w$. Note, that we can express the number of agents N_m^w connected to WAN AP w_m via the set $\{K_k\}_{k\in\mathcal{K}_m}$ as follows:

$$\begin{aligned} N_m^w &= \sum_{k\in\mathcal{K}_m} (M_k^h - K_k)\mathbf{1}_{\left\{M_k^h > K_k\right\}} \\ &= \sum_{r=1}^{N_B} \sum_{\{k\in\mathcal{K}_m | B_m^w(h_k) = b_r^w\}} (M_k^h - K_k)\mathbf{1}_{\left\{M_k^h > K_k\right\}}. \end{aligned} \tag{4.28}$$

[3]Note that the optimal solution to Problem 4.3 might not have this property. However, one can show that for the solution to the original optimization problem 4.1 such property holds.

Let $n_m(r)$ denote the number of sites in S_m^w with WAN rate equal to b_r^w. In the limit $n_m(r) \gg 1$ we can apply the Central Limit Theorem in (4.28), to obtain that:

$$N_m^w = \sum_{r=1}^{N_B} n_m(r)\xi_r\,,$$

where ξ_r is a normally distributed random variable with expectation and variance equal to

$$g(K_r) \triangleq \mathbb{E}\left[(M_r^h - K_r)\mathbf{1}_{\{M_r^h > K_r\}}\right].$$

(Note, that since by our assumption S_k^h have the same sizes for all $k \in \mathcal{K}_m$, we have that M_k^h has the same distribution for all $k \in \mathcal{K}_m$. Hence, $g(K_r)$ depends only on K_r.) Thus, in the limit when $n_m(r) \gg 1$ for $r = 1,\ldots,N_B$, we have that N_m^w is normally distributed with mean and variance equal to:

$$L\left(\{K_k\}_{k\in\mathcal{K}_m}\right) \triangleq \sum_{r=1}^{N_B} n_m(r)g(K_r)\,.$$

Similarly, elaborating on the sum Σ that appears in (4.23), we have:

$$\begin{aligned}
\Sigma &\triangleq \sum_{a_j\in\mathcal{W}_m} \frac{f}{B_m^w(a_j)} = \sum_{k\in\mathcal{K}_m} \frac{f}{B_m^w(h_k)}(M_k^h - K_k)\mathbf{1}_{\{M_k^h > K_k\}} \\
&= \sum_{r=1}^{N_B} \frac{f}{b_r^w} \sum_{\{k\in\mathcal{K}_m | B_m^w(h_k) = b_r^w\}} (M_k^h - K_k)\mathbf{1}_{\{M_k^h > K_k\}} \\
&= \sum_{r=1}^{N_B} \frac{f\,n_m(r)}{b_r^w}\,\xi_r\,.
\end{aligned}$$

Thus, in the limit when $n_m(r) \gg 1$ for $r = 1,\ldots,N_B$, we have that Σ is normally distributed with mean and variance equal to:

$$\sum_{r=1}^{N_B} \frac{f\,n_m(r)}{b_r^w}\,g(K_r)$$

Now, it is simple to see that the constraint (4.25) reduces to requiring that:

$$\sum_{r=1}^{N_B} \frac{f\,n_m(r)}{b_r^w} g(K_r) \le \left(\frac{1}{\theta L\left(\{K_k\}_{k\in\mathcal{K}_m}\right)} + \gamma\right)^{-1} + \varepsilon\,, \tag{4.29}$$

where the variable ε depends on δ and is proportional to the standard deviations of N_m^w and Σ. Note, that the variances of N_m^w and Σ scale as the square root of their respective averages. Thus, when both N_m^w and Σ are large on average, and δ is "moderately small", we can neglect ε

in (4.29). We thus arrive at the following optimization problem:

$$\min \sum_{r=1}^{N_B} n_m(r) \left(\gamma f K_r + \frac{f}{\theta} \mathbf{1}_{\{K_r>0\}} \right).$$

under constraint:

$$\sum_{r=1}^{N_B} \frac{f\, n_m(r)}{b_r^w} g(K_r) \leq \left(\frac{1}{\theta L\left(\{K_k\}_{k\in\mathcal{K}_m}\right)} + \gamma \right)^{-1}. \tag{4.30}$$

It is simpler to treat this problem assuming each K_r takes continuum of values. We might compute $g(K_r)$ replacing M_k^h, for all $k \in \mathcal{K}_m$, by normal random variables η_k, that have the same average and the variance as M_k^h. (This step is supported by the fact that when $\mathbb{E}[M_k^h]$ is large enough the cdf of a Poisson random variable does not differ much at integer points from the cdf of the corresponding normal random variable.) Also, when $\gamma\theta \gg 1$ (delay requirement is not very stringent) we could eliminate the term $\frac{f}{\theta}\mathbf{1}_{\{K_r>0\}}$ from the objective. Then, using Kuhn-Tucker conditions we arrive into the requirement that if $\{K_r^*\}$ is an optimal set of values, then if $K_r^* \neq 0$, we have:

$$\begin{aligned} & n_m(r)\gamma f - \mu \frac{f\, n_m(r)}{b_r^w} g'(K_r^*) \\ & + \mu \frac{\partial}{\partial K_r} \left(\frac{1}{\theta L\left(\{K_k^*\}_{k\in\mathcal{K}_m}\right)} + \gamma \right)^{-1} = 0, \end{aligned} \tag{4.31}$$

and the constant μ is such that the set of $\{K_r^*\}$ obeys the constraint (4.30) with equality. Assuming that $n_m(r)$ for each r is large enough, we can neglect the derivative associated with the last term in (4.31). Then we arrive into a simple requirement for $K_r^* \neq 0$

$$\frac{g'(K_r^*)}{b_r^w} = -\nu, \tag{4.32}$$

where the constant ν is such that is such that the set of $\{K_r^*\}$ obeys the constraint (4.30) with equality. Elaborating on $g_r'(K)$ we get:

$$g'(K_r) = \left(\int_K^\infty (x - K_r) p(x) dx \right)' = -P(\eta_r > K),$$

where $p(\cdot)$ denotes the pdf of the normal random variable with expectation and variance equal to $\mathbb{E}\left[M_k^h\right]$. Combining this with (4.32) yields Proposition 4.3. □

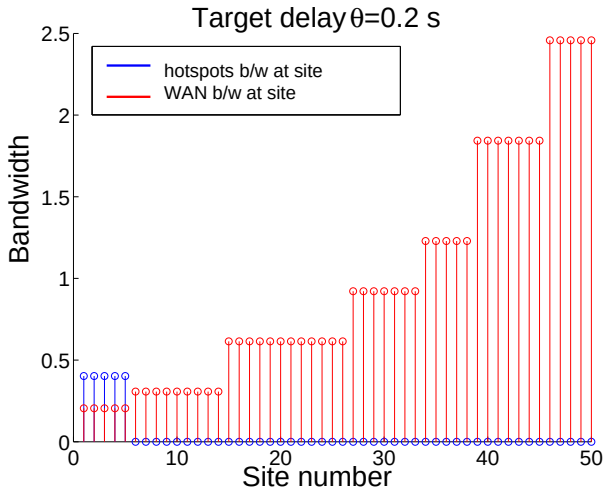


Figure 4.6: Distribution of hotspots backhaul across sites: the red stems show (ordered) WAN rate at each of the 50 sites, the blue stems – the bandwidth for hotspots installed at these sites.

4.4.3 Performance of the optimized system

We implemented the approximate solution of Problem 4.3 given by Proposition 4.3 when the traffic parameters and the geometry of WAN network is as we had for the simulation example in Section 4.4.1. Within a single service zone that has the size of an average typical WAN service zone, we simulated 50 sites with different WAN rate, where rates were generated randomly and independently for each site as described in the setup of Scenario 2. We found that the average delay experienced by agents within the optimized system matches closely the target value, $\theta = 0.2s$.

The results of the optimization for a particular realization of WAN rates, are shown in Figure 4.6. Note that the bandwidth is allocated only to hotspots at the sites that experience the worst WAN rate. We find that on average the total bandwidth required for hotspots in the WAN service zone is less than 2.5Mbs, once appropriate backhaul is allocated to optimally selected sites. For comparison, to guarantee the same average performance in the setup of the Section 4.4.1, where each hotspot was allocated the same bandwidth, one needs about 7.5 Mbps total to be allocated to hotspots on average per service zone of a WAN AP.

4.4.4 Accounting for both spatial density profile and WAN rates profile

Proposition 4.3 can be used to optimize the backhaul bandwidth in a WAN service zone under *homogeneous* user distribution. It is however impractical under non-homogeneous spatial density profiles (see, e.g. Figure 4.1). In that case, clearly, we would like to augment the sites which "see" smaller average WAN rates with more hotspots and associated backhaul, but we would also like to augment the sites with possibly larger WAN rates seeing a larger steady load

users. To optimize the system in this case we need an analogue of Proposition 4.3 that explicitly incorporates the users' spatial density profile in the optimization problem.

In Proposition 4.4 we state the extension of Proposition 4.3 for the case of a non-homogeneous Poisson distribution of users. We will omit its proof since it follows straightforwardly from the proof of Proposition 4.3, however there is an extra scenario assumption which we briefly explain now. We assume that the service zone S_m^w is a union of a few sufficiently large sub-zones, within which the density of users could be approximated via deterministic (constant) values. Now, each of these sub-zones is likely to contain a number of sites S_k^h, $k \in \mathcal{K}_m$, which are identical, in terms of the average WAN rate that is seen from their locations, and the user densities. The realization of the number of users within these sites are i.i.d, conditional on the value of user density in the corresponding subzone, and hence we can use Gaussian approximations to the N_m^w and $\Sigma \triangleq \sum_{a_j \in \mathcal{W}_m} \frac{f}{B_m^w(a_j)}$, similarly to the way we did in the proof of Proposition 4.3.

Let us define:

$$g_k(K_k) = \mathbb{E}[(M_k^h - K_k)\mathbf{1}_{\{M_k^h > K_k\}}], \quad \tilde{L}\left(\{K_k\}_{k\in\mathcal{K}_m}\right) \triangleq \sum_{k\in\mathcal{K}_m} g_k(K_k).$$

Note that the only difference from the corresponding previous definitions is that $g_k(K_k)$ now explicitly depends on the site index, k, since the distribution of the number of users in S_k^h varies with the density of users, λ_k^a. The following Proposition gives the way to optimize backhaul for the service zone with both non-uniform WAN service as well as the non-uniform users' distribution:

Proposition 4.4. *Let the distribution of users within S_m^w be given via a non-homogeneous Poisson process and assume that S_m^w consists of a few non-overlapping regions with constant user's density. Let η_k be a Gaussian random variable, such that $\mathbb{E}[\eta_k] = \mathbf{var}[\eta_k] = \mathbb{E}[(M_k^h - K_k)\mathbf{1}_{\{M_k^h > K_k\}}]$. We have that either $K_k = 0$, or*

$$\frac{\mathbb{P}(\eta_k > K_k)}{B_m^w(h_k)} = \nu,$$

for some constant ν, such that the set $\{K_k\}_{k\in\mathcal{K}_m}$ obeys:

$$\sum_{k\in\mathcal{K}_m} \frac{f}{B_m^w(h_k)} g_k(K_k) = \left(\frac{1}{\theta\tilde{L}\left(\{K_k\}_{k\in\mathcal{K}_m}\right)} + \gamma\right)^{-1}.$$

Example of optimization using spatial density and WAN rates' profiles. Figure 4.7 shows the result of the optimization using Proposition 4.4 for a particular realization of spatial WAN

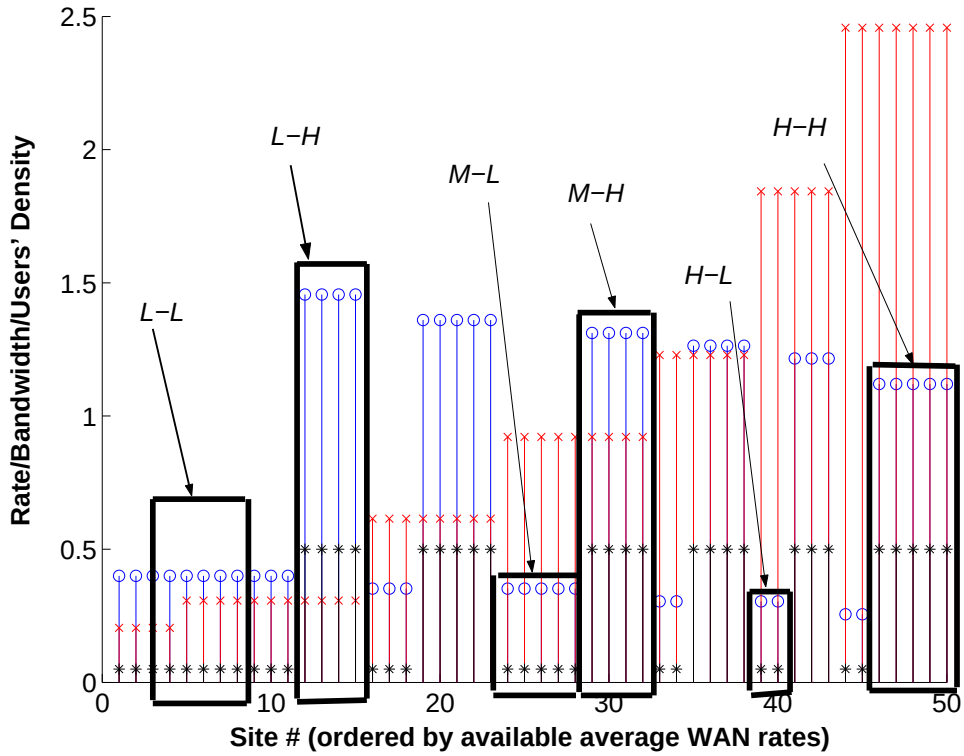


Figure 4.7: Distribution of hotspots backhaul across sites: the red stems show (ordered) WAN rate at each of the 50 sites, the blue stems – the bandwidth for hotspots installed at these sites. The black stems show (scaled) users' density at the corresponding sites (in this experiment $\lambda_k^a \in \{20, 100\}$).

rates profile. The boxes in the figure correspond to classifying various sites according to the available WAN rate and users' density. Thus, for example we marked via $L-H$ the group of sites with relatively "low" average WAN rate and relatively "high" users' density.

In terms of performance we find that Proposition 4.4 yields a design that leads to performance (under congestion-sensitive decision-making) which is quite similar to the target. The difference between the target and the actual performance is, however a bit larger than typically observed with optimal designs according to Proposition 4.3 for the case of spatially homogeneous users' distribution. Thus we conclude, that the effect of neglecting some terms depending on δ when deriving Proposition 4.4 is larger then the effect of similar terms in case with Proposition 4.3. As a general rule of thumb, we suggest to apply Proposition 4.4 with a bit lower target delay, which would lead to some over-provisioning at hotspots, but would still yield large savings in comparison to the over-provisioning for a proximity-based system design.

4.5 Chapter summary

In this chapter we described our first steps towards analyzing a possible future wireless network landscape which incorporates heterogenous technologies, e.g., WAN, LAN, Bluetooth etc. In this setting the criterion used by end devices to select which network to connect becomes an important part of the overall system design. It will not only impact the performance that the user population will see, but also, the resources (e.g., backhaul links density of access points) the providers need to put into place to handle the traffic loads.

To our knowledge this is the first attempt to model and evaluate such heterogeneous systems. In this chapter we showed that congestion dependent decision making is likely to provide on average much better performance to users than simple and more natural proximity based strategies, but only when hotspot bandwidths are limited. Since backhaul links correspond to high recurring costs, it is at this point not unreasonable to expect these to be dimensioned conservatively and thus enabling congestion-dependent decision making by end nodes when presented with WAN and hotspot service options to be worthwhile.

At the same time, we addressed the complementary problem of joint network design for a system incorporating WAN and hotspots to support a spatially distributed set of users. The key insight, is that WAN capacity is particularly valuable, because it permits statistical multiplexing of spatial fluctuations in user loads over a wide area. By contrast hotspots have the potential to substantially and inexpensively enhance capacity in a restricted area. Thus under the uniform loads, as we show in this chapter, it is the case that WAN resources are typically used as much as possible with only the necessary bandwidth allocated across hotspots to alleviate overloads on the WAN. However, if there are spatial inhomogeneities in the capacity the WAN can provide to users, or in the characteristics of the load, the synergies between these technologies may take a different form. Indeed one may conclude that hotspots and the associated backhaul bandwidth is truly worthwhile at particular spatial locations with a high steady (i.e., low variance) offered load and where the WAN is not able to provide reasonable service. This chapter shows that a joint system design is likely to exploit such variations in order to reduce overall system cost significantly.

Chapter 5

Interference-mitigating load balancing in "tightly coupled" multi-provider wireless networks

5.1 Introduction

In this chapter we will consider the design of heterogeneous wireless systems in which decision-making strategies that guide the users' assignment to the WAN or hotspots incorporate some extra information about the current system state, which might not be available locally for all users. For example, the additional information may include the utilization of all hotspot APs falling within a WAN service zone containing the user – this is different from the setup of Chapter 4 where a user's agent was assumed to only be able to estimate the congestion state of the WAN and the hotspot APs that are located within some limited vicinity of the user. Such extra information might be aggregated at the RAC (see Chapter 4) and then signalled to users' agents, which in turn would make more favorable, from the point of view of the overall network, decisions regarding the choices of APs.

Note that providing additional feedback requires that the communication protocols at the WAN and hotsptots have to be augmented in order to include extra signaling/information exchange between the APs, users and (possibly) a separate networking entity, e.g. RAC (see Chapter 4). Clearly the interaction between the users and the network is potentially more complex for such "tightly coupled" heterogeneous networks, and enabling such interaction is likely to come at the cost of additional network complexity. However, the extra complexity is likely to payoff by enabling tighter control of network performance, leading to more efficient resource utilization and translating to further performance gains.

The purpose of this chapter is to formulate and evaluate *centralized and distributed* decision-making algorithms that could be implemented within the context of a "tightly cou-

pled" wireless data network. These algorithms lead to gains in performance due to the efficient redistribution of load between the WAN and WLAN networks. We refer to these gains as *"load balancing" gains*, since the study presented in this chapter is related to the large body of work on load balancing in distributed computer systems and wireline networks, see e.g. [67]–[51]. Application of methods developed therein is, however, often challenged by the specifics of wireless applications. For example, notions of resource and link capacities are not as straightforward to define in a wireless setting as in the wireline case. The quality of wireless channels suffers from fading and interference and thus the physical data rates/service rates that are available at users' spatial locations often significantly depend on system state.

Research targeting load-balancing in the wireless setting is still in its infancy. In [44] two threshold-based algorithms are evaluated, which bias users that wish to download small files or are moving too fast to select WAN APs. It is shown that selecting the thresholds appropriately in some scenarios leads to better load balancing among the layers. However, these assignment decisions are oblivious to APs' utilization, thus there is no guarantee that the system operation point is optimal.

In this chapter we will derive load-balancing decision metrics that tie together the average physical channel rates available at users' locations, proximity to access points, utilization of resources and current demand for particular services. The most important feature that distinguishes this work from, e.g. [44], is how we incorporate the interference at the WAN layer as a factor biasing the load-balancing decisions. As in Chapter 4, our model is applicable to systems where no power control takes place on the downlink at the WAN layer: a key example of such a system is Qualcomm's 1XEV-DO [55]. In 1XEV-DO users are "roughly" served in a "generalized processor-sharing" fashion [56], and only one user is served per time slot with full power allocated at the WAN AP. Thus in such systems the larger the fraction of users routed to a particular WAN AP the larger the fraction of time that the WAN AP has to be active on average. This in turn makes the WAN AP create more interference to its neighboring WAN APs, and forces degradation in the service quality seen by users served by neighboring WAN APs. We will show that algorithms which factor such interference at the WAN layer lead to improved performance, especially in systems with spatially asymmetric loads.

It is worth mentioning that the problem of "other-cell" interference is usually also present at WLAN layer. In this study we chose to neglect this problem for the following reasons. First, we assume that WLANs are spatially spaced, so that the interference a WLAN creates at service zones of peer WLANs can be viewed as negligible. Thus, for example, access points in IEEE 802.11 systems operating in the unlicensed band, have to obey rigid FCC regulations in the maximum power they use, which implies that the signal emitted by such access points is only strong enough within the limited vicinity of the access points. Secondly, in most practical scenarios even when interference is present, the physical rates for communication with the

WLAN are mostly limited not by the wireless access bandwidth but by the available bandwidth provided at the backhaul [63]. Thus for example, although WiFi access points can operate at physical rates of over 10Mbps, it is rarely the case that the available backhaul bandwidth exceeds 1Mbps: bachhaul capacity usually incurs high recurring costs.

In this chapter we will reuse the geometric models for service zones of the WAN and WLAN APs which we built in Chapter 2. In Sections 5.3 and 5.4 we will derive our algorithms for a generic geometric model obeying the set of Assumptions 2.1. For most of our simulations we will, however, use a geometric setup in which WAN and WLAN APs are placed on regular meshes and in which the service zones are defined as described in Section 2.3.2. Also, as we move from a "per WAN service zone" analysis in Chapter 4 to an analysis of cell interactions within a multi-cell heterogeneous wireless system, we will need to incorporate some additional queueing modeling and approximations.

The outline of this chapter is as follows. We recapitulate the additional assumptions regarding the interference model, etc. in Section 5.2. In Section 5.3, under the assumption that other-cell interference at the WAN layer is negligible, we derive decision-metrics that are used as a basis to formulate our load-balancing algorithms. In Section 5.4 we describe how one can adjust the decision-metrics so as to incorporate the "cost" of other-cell interference at the WAN. The adjusted algorithms perform equally well in scenarios with both significant and small other-cell interference. We demonstrate this in Section 5.5 where we report the results of our simulations.

5.2 Additional assumptions

5.2.1 Traffic model

We will concentrate on the so-called semi-dynamic [69] scenario. The setup is very similar to the one we had in Chapter 4: mobiles generate requests for file transfers, that arrive at random spatial locations, and stay at these locations for the duration of the file transfer. The arrival of requests is described by a stationary, possibly nonhomogeneous, spatial Poisson process, i.e. the numbers of arrivals per unit time within disjoint spatial regions are independent, and the number of arrivals per unit time for a region ΔS is Poisson distributed with parameter given by $\int_{y\in\Delta S}\lambda(y)dy$. We also let the file sizes associated with the requests to be independent and generally distributed with, possibly, spatially-dependent means, denoted by $f(y)$, for $y \in D$.

5.2.2 Service type at APs

We shall again postulate that APs serve queued requests in a *processor-sharing* fashion and no request is blocked from service. In our simulation models the time is divided into slots of

duration 1.67 ms and within each slot a small portion of the requested file transfer for a single user is realized. To derive our algorithms we will assume that users with active requests are served in a round-robin fashion which neglects the gains from multi-user diversity, exploited, for example in the HDR protocol [55]. This assumption, will greatly simplify our analysis and will be relaxed when evaluating the proposed algorithms via simulation.

5.2.3 Data rates

Each time slot the amount of data that is served to a user depends on the instantaneous data rate, available at the user's location. Note that there are two factors that limit this rate: the received signal quality and the bandwidth of the backhaul of the corresponding APs. Our numerical experiments have been performed under the assumption that the backhaul at WAN APs matches the channel rates, while backhaul capacity for WLAN APs is very limited.

In our simulations, we will let user's physical data rate on each time slot be a function of instantaneous value of the SINR only. (An example of such SINR-to-rate mapping that we use to compute data rates at WAN layer is provided in Figure 4.5.) Recall that our focus is on systems with no down-link power control, thus whenever APs are active they use the same power level. Let $\mathcal{M}(t)$ denote the set of WAN APs' transmitting during time slot t in the same frequency band as WAN AP w_1. Then the SINR $\eta^w(x,t)$ for a user located at $x \in S_1^w$ which is served at slot t by WAN AP w_1 can be expressed as:

$$\eta^w(x,t) = \frac{P_1^w L_1^w(x,t)}{N(x) + \sum_{m \neq 1} \mathbf{1}(m \in \mathcal{M}(t)) P_m^w L_m^w(x,t)}, \tag{5.1}$$

where P_n^w and $L_n^w(x,t)$ for $n = 1, \ldots, M$ denote, respectively, the power levels that are used by the WAN APs and the attenuation factors at location x at time t, for signals emitted at WAN AP, and $N(x)$ denotes the ambient noise power at location x.

5.2.4 Attenuation

Generally[1] we will model the attenuation from each AP as a superposition of a large scale path loss and, independent slow (log-normal) and fast (Rayleigh) fading components. Thus at time slot t, $L_n^w(x,t)$ is independently drawn from an exponential distribution with time-independent mean $\bar{L}_n^w(x)$ given by:

$$\bar{L}_n^w(x)(dB) = PL(w_n,x)(dB) + Z_\sigma,$$

where $PL(w_n,x)$ denotes the large scale path loss component and Z_σ a zero-mean, normally distributed random variable with standard deviation σ. For the large scale path loss we adopt

[1] In some simulations we will "turn off" several fading components to better illustrate the concepts we develop.

(from [70], p. 642) the formula[2]:

$$PL(w_n,x)(dB) = -118 - 10\alpha \log_{10}(|x-w_n|)\,, \tag{5.2}$$

where α is the path loss exponent and the distance $|x-w_n|$ is expressed in miles. In our simulations we use the range $2-4$ for α, and the range $6-10$ for σ.

5.2.5 System objective

As in Chapter 4 we will denote by τ the decision-making control strategy, i.e. the rule that determines to which AP an incoming request is routed. We will evaluate the efficiency of strategy τ on the basis of how well this strategy optimizes a certain system objective, given a stationary spatial load of requests for file transfers. As we mentioned already in Chapter 4 when the WAN service is severely non-uniform using worst user's performance as a performance metric might lead to designs with unreasonable and inefficient usage of network resources. Thus in this chapter we change the performance metric from the "worst users' performance" to the "average across users performance", given by (4.6). Our motivation for this change comes from both the fact that we would be like to consider also severely non-uniform WAN coverage scenarios, as well as to be able to simplify our derivations. Thus, we set the system objective to the the total mean queueing backlog within the system:

$$U_{system}(\tau) = \sum_{m=1}^{M} \mathbb{E}^{\tau}[Q_m^w] + \sum_{k=1}^{K} \mathbb{E}^{\tau}[Q_k^h]\,, \tag{5.3}$$

where $\mathbb{E}^{\tau}[Q_m^w]$ and $\mathbb{E}^{\tau}[Q_k^h]$ denote the average queueing backlog at WAN AP w_m and hotspot AP h_k under the decision-making policy τ. In expression (5.3) we use the convention that $U_{system}(\tau) = \infty$ if the strategy τ results in unstable queues. Note that whenever the system is stable under τ, minimizing $U_{system}(\tau)$ also corresponds (via Little's law) to minimizing the *average delay* experienced by a typical user in the system, and thus results in the same optimization as under the performance metric given by (4.6).

5.3 Load balancing when other-cell interference is negligible

As reflected by equation (5.1), the instantaneous SINR seen by users connecting to the WAN depends, potentially, on the activity pattern of all WAN APs in the system. This leads, potentially, to coupling between the queue states of all WAN APs in the system and the service rates that are seen by the users. The analysis of this coupling in general is quite difficult. We postpone

[2]For small $|x-w_n|$ this is a poor model. We simply put $L_n^w(x,t) = 0$ if the expression yields a positive value.

discussion of this general case to Section 5.4.

Instead, in this section we will start by formulating load-balancing algorithms for scenarios where the other-cell interference at the WAN layer is small enough that it can be neglected. Such scenarios arise, in, e.g., systems where spectrum is spatially divided between different WAN APs so that no two adjacent service zones share the same frequency band. Our approach in this section also enables us to roughly model scenarios where all WAN APs use the same spectrum, but the attenuation is high and only a small fraction of users at service zone boundaries are substantially affected by other-cell interference.

By Assumption 2.1.3 two different requests that arrive within the service zone of the same WLAN AP can only be routed to the WLAN AP or the *same* WAN AP. When no other-cell interference is present at both WAN and WLAN levels, the physical data rates are only affected by fading. It follows that, under any decision-making strategy, the service rates for users within service zones of different WAN APs are independent. Thus the set of optimal rules includes one τ which is decomposable into a collection $\{\tau_m\}_{m=1}^{M}$, where τ_m for $m = 1,\ldots,M$ are decision strategies operating independently on requests originating within S_m^w, for $m = 1,\ldots,M$ respectively. We will focus on designing each τ_m separately, and without loss of generality we concentrate on τ_1.

Let $\mathcal{K}_m$ denote the indices of WLAN APs that fall within the service zone S_m^w. Taking into account the above discussion, the optimization reduces to finding a decision strategy τ_1 that minimizes the objective:

$$U_1(\tau_1) = \mathbb{E}^{\tau_1}[Q_1^w] + \sum_{k\in\mathcal{K}_1} \mathbb{E}^{\tau_1}[Q_k^h],$$

over some appropriately defined set of strategies.

5.3.1 Centralized adaptive load balancing.

We first consider a family of strategies where incoming requests are routed to APs in a probabilistic fashion. We will assume that the decision-making entity is able to differentiate between a finite number of disjoint service classes, and associates the incoming requests with a particular class based on the average physical data rates available at the requests' locations. In other words, each request is associated with a given class which in turn quantifies the mean rate that the request can achieve to the WAN and WLAN. Our goal is to optimize system performance by selecting appropriate per-class routing probabilities.

We note that the results of this subsection are not particularly new – they combine classical results in adaptive routing for wireline networks [71] with recent results [56] using generalized processor-sharing queueing disciplines to modeling "proportionally fair" scheduling algorithms in wireless networks.

Let us denote by $B^w(x)$ the time average physical data rate available at $x \in S_1^w$ from WAN

AP w_1. Similarly, let $B^h(x)$ denote the time average of the physical data rate available at $x \in S_k^h$, from WLAN AP $h_k \in S_1^w$. Based on this pair of rates the request at location x is assigned to one of the disjoint classes, which we will denote via $\mathcal{B}_i$, for $i = 1, \ldots, I$. We will assume that the probability of routing a request that belongs to class $\mathcal{B}_i$ to the WAN AP w_1 is given by P_i and let $\mathbf{P} \equiv \{P_i\}_{i=1}^I$.

Using the results in [56], we may express the system objective in terms of the vector $\mathbf{P}$ as:

$$U_1(\mathbf{P}) = \frac{\rho_1^w(\mathbf{P})}{1-\rho_1^w(\mathbf{P})} + \sum_{k \in \mathcal{K}_1} \frac{\rho_k^h(\mathbf{P})}{1-\rho_k^h(\mathbf{P})}, \tag{5.4}$$

where the utilization $\rho_1^w(\mathbf{P})$ of the WAN AP w_1 is given by:

$$\rho_1^w(\mathbf{P}) = \int_{y \in \bar{C}_1} \frac{\gamma(y)\mathbf{1}(y \in \mathcal{B}_i)}{B^w(y)}\, dy + \sum_i^I P_i \int_{y \in C_1} \frac{\gamma(y)\mathbf{1}(y \in \mathcal{B}_i)}{B^w(y)}\, dy, \tag{5.5}$$

and the utilization $\rho_k^h(\mathbf{P})$ of WLAN AP $h_k \in S_1^w$ is given by:

$$\rho_k^h(\mathbf{P}) = \sum_i^I (1-P_i) \int_{y \in S_k^h} \frac{\gamma(y)\mathbf{1}(y \in \mathcal{B}_i)}{B^h(y)}\, dy. \tag{5.6}$$

In the above expressions $\gamma(y) \equiv f(y)\lambda(y)$, and we denote by C_1 the set of locations in S_1^w covered by some WLAN (we use $\bar{C}_1$ for locations not covered by any WLAN), i.e. $C_1 = \bigcup_{k \in \mathcal{K}_1} S_k^h$ ($\bar{C}_1 = S_1^w \setminus C_1$). Our optimization problem is then given by:

Problem 5.1.

$$\min_{\mathbf{P}} \{U_1(\mathbf{P}) \mid 0 \leq \mathbf{P} \leq 1\}$$

The following proposition establishes the convexity property of this cost function, as is the case for average delay in Jackson networks [71]. For completeness, we include a proof addressing the specifics of our setup.

Proposition 5.1. *Problem 5.1 is convex. The gradient elements of $U_1(\mathbf{P})$ are given by:*

$$G_i \equiv \frac{\partial U_1(\mathbf{P})}{\partial P_i} = G_i^w - G_i^h, \tag{5.7}$$

where

$$G_i^w = \frac{\Gamma_i^w}{(1-\rho_1^w)^2}, \qquad G_i^h = \sum_{k \in \mathcal{K}_1} \frac{\Gamma_i^h(k)}{(1-\rho_k^h)^2}, \tag{5.8}$$

and Γ_i^w, $\Gamma_i^h(k)$ *are defined as:*

$$\Gamma_i^w = \int_{y\in C_1} \frac{\gamma(y)\mathbf{1}(y\in\mathcal{B}_i)}{B^w(y)}\,dy, \quad \Gamma_i^h(k) = \int_{y\in S_k^h} \frac{\gamma(y)\mathbf{1}(y\in\mathcal{B}_i)}{B^h(y)}\,dy,$$

Proof. It is straightforward to verify the expression for the gradient of the objective. The convexity in $\mathbf{P}$ is also easy to check by showing that the Hessian $\mathbf{H} = [H_{i,j}] \equiv \left[\frac{\partial^2 U_1}{\partial p_i \partial p_j}\right]$ is nonnegative-definite. Indeed, observe that:

$$H_{i,j} = \frac{2\Gamma_i^w\Gamma_j^w}{(1-\rho_1^w)^3} + \sum_{k\in\mathcal{K}_1} \frac{2\Gamma_i^h(k)\Gamma_j^h(k)}{(1-\rho_k^h)^3}.$$

Thus for any real vector with components $\{e_i\}_{i=1}^I$ we have:

$$\sum_{i,j=1}^{L} H_{ij}e_ie_j = \frac{2}{(1-\rho_1^w)^3}\left(\sum_i^I \Gamma_i^w e_i\right)^2 + \sum_{k\in\mathcal{K}_1}\frac{2}{(1-\rho_k^h)^3}\left(\sum_{i=1}^I \Gamma_i^h(k)e_i\right)^2 \geq 0.$$

The convexity of Problem 5.1 follows now from the convexity of the optimization region. □

Since Problem 5.1 is convex, any version of a gradient descent method could be used to obtain globally optimal per-class routing probabilities $\mathbf{P}$. For example we can iteratively adapt the routing probabilities in a "greedy" fashion according to:

$$\mathbf{P}(t+1) = \min[1, [\mathbf{P}(t) - a\nabla U_1(t)]^+], \tag{5.9}$$

where $a > 0$ is some appropriately chosen constant, and the gradient $\nabla U_1(t)$ is estimated at each time step, t, using Proposition 5.1.

We can now describe an implementation of the centralized adaptive load-balancing algorithm based on gradient estimation. In our implementation, WAN and WLAN APs are able to obtain initial estimates for the average physical data rates at locations of all incoming requests. Such estimates could in practice be obtained for each dual-mode device during initial session set-up and then maintained throughout sessions' life-time. The incoming requests are first routed to the central controllers, residing at each WAN AP, and the controllers forward the request to either WAN or WLAN AP according to the current value of the routing probabilities. The controller at, e.g. WAN AP w_1 is able to communicate (via a wired connection) to all WLAN APs in its service region S_1^w and thus maintains a data base that includes smoothed estimates for utilizations of the WAN and WLAN APs within S_1^w (equivalently, fraction of time each AP is busy), estimates for the average data rates (at both WAN and WLAN layers) for each active mobile in the system, and current values of the routing probabilities.

Whenever a request is forwarded to the controller, the information on the average data rates is used to associate request with a particular class of service. The size of file F and the corresponding estimates $\hat{B}^w$ and $\hat{B}^h$ of average physical data rates at WAN AP w_1 and WLAN AP $h_k \in S_1^w$, associated with request of class $\mathcal{B}_i$, are then used to update the values of Γ_i^w and $\Gamma_i^h(k)$ – we simply keep a finite history of the entries given by $F/\hat{B}^w$ and $F/\hat{B}^h$ and compute a running average of these entries. Given the estimates for utilization of the WAN AP ρ_1^w, utilization of WLAN APs ρ_k^h, $k \in \mathcal{K}_1$, estimates of Γ_i^w, $\Gamma_i^h(k)$, for $i = 1,\ldots,I$, $k \in \mathcal{K}_1$, the controller at w_1 updates the estimate of the gradient via (5.7,5.8), and the values of per-class routing probabilities, via (5.9).

5.3.2 Design tradeoffs

In an ideal scenario, where the estimations of various quantities are perfect, the adaptation of routing probabilities is slow enough and the arrivals to the system are stationary, the centralized algorithm eventually yields a set of optimal per-class routing probabilities. In practice, however, various quantities will be computed up to a certain precision, yielding errors in gradient estimation. Iterative stochastic approximation methods [72] can, in principle, be applied to cope with such errors. However, most of these methods are formulated assuming various system's parameters and inputs (e.g. request arrivals) are stationary, and make the adaptation rate (controlled by constant a in (5.9)) to be gradually diminishing to zero. Note that, request arrivals in realistic scenarios are rarely stationary, meaning that the adaptation rate has to be large enough so as the algorithm reacts in a timely fashion to changes in load distribution.

In Figure 5.2 we demonstrate the effect of the design tradeoff explained above, for a particular simulation experiment. Figure 5.1 shows the geometric setup: we have an isolated WAN AP operating in the area, whereas in addition, part of the area is covered by a few WLANs. WLANs support data rates of exactly 600 kbps (WLAN rates are limited by the backhaul bandwidth), while WAN physical rates within the region span the range $\mathcal{R} \equiv [0, 2458]$ kpbs and have a spatial average of about 1650 kbps. The range of rates $\mathcal{R}$ is partitioned among I classes, and we assign a request to class i, if its corresponding data rate, falls within the interval $[2458(i-1)/I, 2458i/I]$, for $i = 1, 2\ldots, I$.

In each of the three experiments the total number of classes, $I = 10$, and the random number generator is initialized with a common value. We vary the adaptation rate a, but keep the smoothing windows in estimating average utilizations and traffic demands across classes fixed. Note that the case with $a = 0.01$ results in a non-converging dynamics of routing probabilities, whereas the values of probabilities forever oscillate around the optimal values. At the same time, we note that the dynamics of WAN utilization is virtually identical in all three cases, which eventually stipulates that the mean delays for a typical request under adaptations with various values of a are very similar. In fact, we have experimentally validated that, although

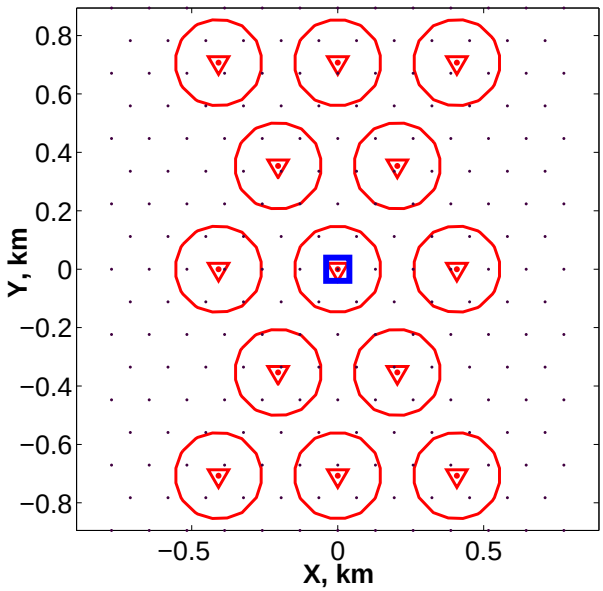


Figure 5.1: A rectangular region served by a single WAN AP, indicated by the box in the center and a few WLANs, shown as triangles. WLAN service areas coincide with the discs, while the WAN AP services the whole region.

larger values of adaptation rates do result in oscillating dynamics for routing probabilities, the overall average performance of the system over time is quite similar or even better than the one for slower adaption rates.

Another design tradeoff is associated with the granularity of partitioning into rate classes, $\{\mathcal{B}_i\}_{i=1}^{I}$ which by itself represents a design challenge. Finer granularity leads, potentially, to decisions per class being better "tuned" to the traffic characteristics of each class representative, (as opposed to being tuned to the characteristics of a "typical" class representative, for coarse classes partitioning) which in turn potentially leads to improved performance at the optimal operating point. At the same time, a large number of classes might be computationally burdensome to handle; and in addition, the effect of errors in estimating Γ_i^w and $\Gamma_k^h(i)$, for $i = 1, \ldots, I$ is likely to increase as the number of classes grows, since the arrivals per class become more rare.

5.3.3 Efficient distributed heuristic for load balancing

Implementation of the centralized controller requires tight coordination between the WAN and WLANs. In practice it is desirable not to require such coordination: WAN and WLAN networks could be operated by different providers and may not even be aware of each other's existence. To secure good performance for such networks, the intelligence has to be moved to the dual-mode devices themselves and implemented via software "agents", which select suitable access points with minimal feedback from WAN/WLAN APs. In this subsection we provide a particular

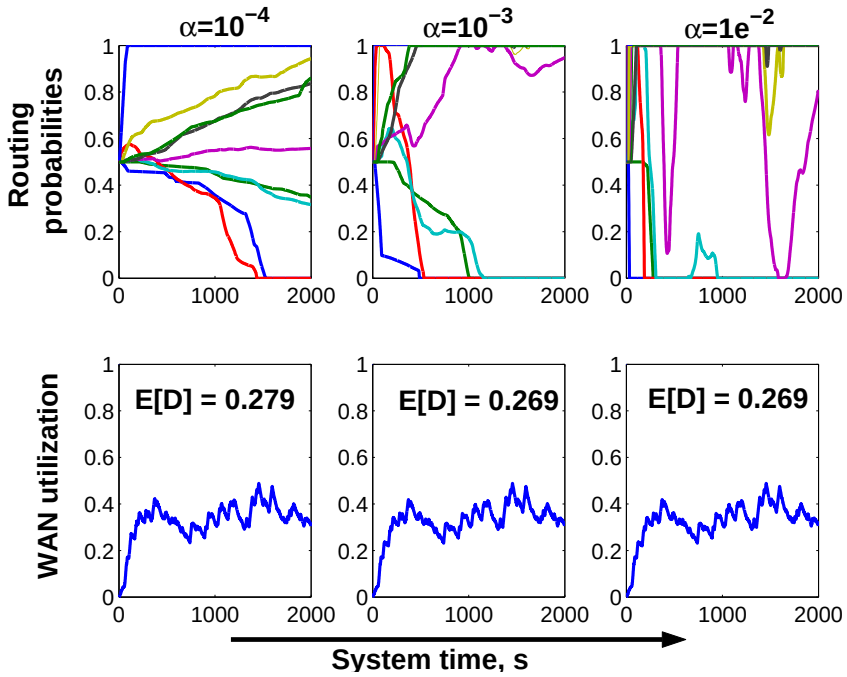


Figure 5.2: Dynamics of per-class routing probabilities and WAN utilization vs. adaptation rate, a.

design for such agents along with the design of the APs' feedback.

Unfortunately, the computation of gradients given via Proposition 5.1 is not amenable to distributed implementation. To overcome this problem, we reformulate our optimization problem by switching to more convenient variables. We partition S_1^w into a large number L of disjoint sets which we denote $\Delta S_1, \Delta S_2, \ldots, \Delta S_L$, containing the representative locations $y_1, y_2, \ldots, y_L$ respectively. When L is sufficiently large, one can approximate, for $y \in \Delta S_i$: $p(y) = p(y_i) \equiv p_i$, $\gamma(y) = \gamma(y_i)$, $B^w(y) = B^w(y_i)$, and $B^h(y) = B^h(y_i)$. Let $\mathbf{p}$ denote the vector $\{p_i\}_{i=1}^L$, then we obtain the following representation of the system objective as a function of vector $\mathbf{p}$:

$$U_0(\mathbf{p}) = \frac{\rho_1^w(\mathbf{p})}{1-\rho_1^w(\mathbf{p})} + \sum_{k \in \mathcal{K}_1} \frac{\rho_k^h(\mathbf{p})}{1-\rho_k^h(\mathbf{p})},$$

where

$$\rho_1^w(\mathbf{p}) = \sum_{i=1}^{L} \frac{\gamma_i |\Delta S_i| p_i}{B^w(y_i)}$$

$$\rho_k^h(\mathbf{p}) = \sum_{i=1}^{L} \mathbf{1}(y_i \in S_k^h) \frac{\gamma_i |\Delta S_i| (1-p_i)}{B^h(y_i)},$$

and $|\Delta S_i|$ denotes the area of ΔS_i. The corresponding optimization problem could be defined similarly to Problem 5.1, but for the function $U_1(\mathbf{p})$ in place of $U_1(\mathbf{P})$.

Problem 5.2.

$$\min\{U_1(\mathbf{p})|\ 0 \leq \mathbf{p} \leq 1\}\,.$$

We also can state an analogue of Proposition 5.1 as follows:

Proposition 5.2. *Problem 5.2 is convex. The gradient elements of* $U_1(\mathbf{p})$ *are given as:*

$$\frac{\partial U_1(\mathbf{p})}{\partial p_i} = \gamma_i |\Delta S_i| \left(G_i^w - G_i^h\right), \tag{5.10}$$

where

$$G_i^w = \frac{1}{B^w(y_i)(1-\rho_1^w)^2}\,,$$
$$G_i^h = \sum_{k \in \mathcal{K}_1} \frac{\mathbf{1}(y_i \in S_k^h)}{B^h(y_i)(1-\rho_k^h)^2}\,. \tag{5.11}$$

Based on Proposition 5.2 we suggest the following decision-making strategy. As before, prior to generating any requests, each dual-mode device has to establish a connection with the APs which contain the device in their service zones. During initial session setup and within the session lifetime, the devices maintain information about the average physical data rates that are available for communication with nearby APs. Prior to requesting a file download from one of the APs, a device at location $y_i \in S_1^w \cap S_k^h$ asks both WAN AP w_1 and WLAN AP h_k to signal their current measured utilizations ρ_1^w and ρ_k^h. The device then computes an estimate $\hat{G}_i^w$ and $\hat{G}_i^h$ via (5.11) using the estimates for the respective utilizations and average physical data rates. It forwards the download request to the WAN AP w_1 if $\hat{G}_i^w - \hat{G}_i^h < 0$ and addresses the request to WLAN AP h_k otherwise.

The proposed approach is based on the observation that the routing probability p_i has to be decreased if the derivative $\frac{\partial U_1(\mathbf{p})}{\partial p_i}$ is positive. By imposing "hard" decisions on the agents we diminish the per unit time average of the number of requests incoming to the WAN AP from the set of locations around $\Delta S(y_i)$. The distributed algorithm is heuristically motivated by Proposition 5.2. The decisions that agents make are no longer probabilistic and thus, strictly speaking, the analysis based on the assumption that arrivals to APs follow Poisson processes no longer holds.

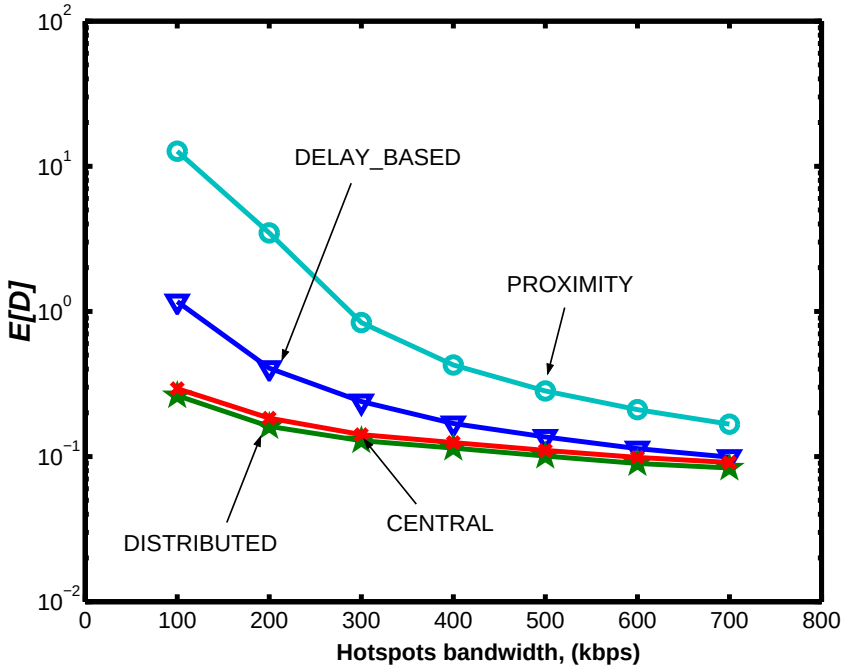


Figure 5.3: Performance of centralized and distributed load balancing algorithms vs. greedy delay-based and proximity-based decision-making strategies.

5.3.4 Implementation of distributed heuristic within a "loosely coupled" heterogeneous network

Note, that under distributed decision-making each agent needs to get a feedback only from the WAN and hotspot APs that contain it in their service zones. Thus, potentially, the distributed decision-making in the sense of Proposition 5.2 could even be implemented in the framework of a "loosely coupled" network, provided each agent is able to estimate the utilization of the WAN and a hotspot APs in its vicinity. To estimate an AP's utilization each agent can either send probing "ping" packets or use the history of download delays. It is a well known [56] property of generalized processor-sharing queueing disciplines, that (when there is no blocking) the expected delay of a file transfer $\mathbb{E}\left[D \mid \frac{f}{B} = b\right]$, conditioned on a particular value of a normalized service requirement[3] $\frac{f}{B^w}$ is given by:

$$\mathbb{E}\left[D \mid \frac{f}{B^w} = b\right] = \frac{b}{(1-\rho)}.$$

By sending ping packets of size f, maintaining the estimate of average bit-rate $\hat{B}_i^w(a_i)$ and $\hat{B}^h(a_i)$ and observing the delays of $\hat{D}^w$ and $\hat{D}^h$ of the ping packets when sent towards WAN and hotspot APs, an agent a_i would be able to obtain estimates $\hat{\rho}^w$ and $\hat{\rho}^h$ of utilizations of both

[3]Normalized service requirement is defined as the ratio of the file size to the average service rate available at the agent's location.

WAN and hotspot AP. Indeed, we have:

$$\hat{\rho}^w = 1 - \frac{f}{\hat{D}^w \hat{B}^w}, \qquad \hat{\rho}^h = 1 - \frac{f}{\hat{D}^h \hat{B}^h}.$$

With these estimates, Proposition 5.3.3 could be used straightforwardly to make decisions contingent on comparing the values of G_i^w and G_i^h. Different variations on this scheme are also possible, e.g. each device can maintain a history of estimates $\hat{\rho}^w$ and use smoothed (via, e.g. exponential filter) estimates of $\hat{\rho}^w$ and $\hat{\rho}^h$ in its decision-making.

5.3.5 Performance of interference-unaware load balancing algorithms

Apart from the fact that the distributed version of the load balancing algorithm does not require a centralized controller to make routing decisions, it has another attractive feature. This feature is represented by the fact that there is no need to estimate the current traffic demand at various locations, or on a per-class basis as done by centralized implementation when estimating the quantities Γ_i^w and $\Gamma_i^w(k)$. As mentioned earlier, the imprecision of these quantities only contributes to errors in estimating of gradient of the objective.

For most of our simulation experiments we found that distributed decision-making strategy performs either as well or even better than the centralized one. In Figure 5.3 we illustrate this by comparing the load-balancing performance that is achieved in simulation model described in Section 5.2 by deploying either a centralized or distributed strategy. The mean delay seen by a typical request is plotted vs the backhaul bandwidth, available at each WLAN. To provide a benchmark, as in Chapter 4 we also use a simple and more natural (at least in current applications) proximity-based routing strategy, under which all requests that emerge within a service zone of a WLAN are simply routed to the corresponding WLAN AP.

On Figure 5.3 we also show the performance of load-balancing algorithms to the performance of a greedy congestion-sensitive decision-making of Chapter 4. Under the latter strategy each device simply sends to the AP which grants it the least file-transfer delay, which it deduces either from previous download history and/or by periodically sending short "ping" messages. Recall, that greedy delay-based decision-making is optimal only when the WAN service is spatially uniform, and performance metric is given by the worst users average delay. It is thus clear why both centralized and distributed load-balancing strategies significantly outperform the strategies based on greedy delay minimization. At the same time we note, that greedy delay-based decision-making still enables a large percent of the total gains in performance of load-balancing strategies over proximity based decision-making.

Finally, Figure 5.4 compares the performance of a distributed strategy realized in a tightly coupled heterogeneous network with the performance of distributed decision-making realized within a loosely-coupled network, which we described in Section 5.3.4. We note that

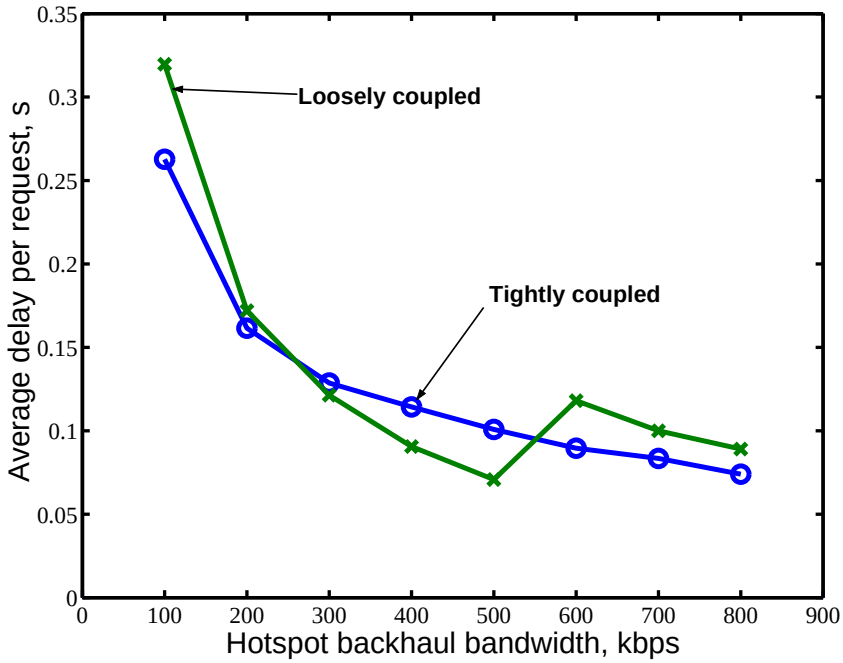


Figure 5.4: Performance of distributed algorithms realized within a tightly and a loosely coupled heterogeneous networks.

the two graphs are quite close to each other. This, however does not mean that there is no benefit in introducing tight coupling – this just indicates that distributed decision-making within both tightly and loosely coupled contexts is likely to perform similarly. In the next section we will, indeed, demonstrate the value of tight coupling by considering "interference-aware" algorithms.

5.4 Interference-aware load balancing

In this section we return to a more general case, in which the activity pattern of a WAN AP can have a substantial impact on the interference seen within service zones of adjacent WAN APs. Straightforward application of the interference-unaware distributed or centralized decision-making rules, described in Section 5.3 might result in poor overall system performance. We start this section by explicitly constructing a simulation scenario that captures such circumstances.

5.4.1 Motivating example

Figure 5.4 shows the geometry for our experimental setup. We have a region that is served by a mesh of WLANs, whereas each WLAN uses enough power to cover its respective Voronoi region. There are also six WAN APs that share a service zone boundary with the WAN AP w_1 located at the center of the region. The experiment that we describe models an asymmetric

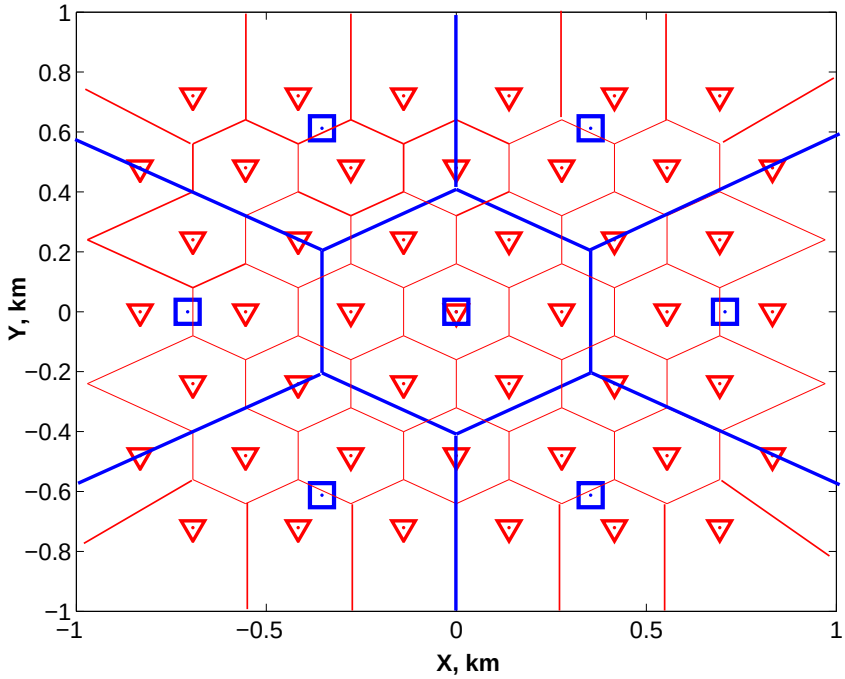


Figure 5.5: Geometric setup: WAN APs shown as boxes and WLAN APs are shown as triangles. The power levels at WLANs are sufficient to cover their service zones represented by corresponding Voronoi cells. The power levels used at WAN APs are large enough to ensure "interference-limited" regime of operation.

traffic scenario, in which the aggregate load within S_1^w greatly exceeds the load for the adjacent WAN service zones. To create even more load asymmetry we assume that WLANs falling within S_1^w are *shut off*, in which case the WAN AP w_1 has to serve all requests emerging within S_1^w.

We summarize the particular traffic characteristics and other assumptions on system parameters in Table 5.4.1. To exclude the randomness induced by log-normal and Rayleigh fading, in this experiment we will only model large scale attenuation, given by equation (5.2), where the value of path loss exponent, $\alpha = 2.0$. Clearly, $\alpha = 2.0$ can hardly correspond to any realistic scenario[4], but is chosen to maximize the effect of other-cell interference on performance so we can better illustrate our ideas. We also let the power levels at APs to be the same, and large enough, so that the effect of ambient noise in equation (5.1) is negligible i.e., the system operates in the so-called "interference-limited" regime.

Figure 5.4.1 shows simulation results for this scenario where we vary the file arrival rate within the service zone of the WAN AP in the center. The results are interesting in that they demonstrate how the proximity-based strategy, which is probably the worst possible decision-making strategy under low utilizations, becomes desirable with increased load at central WAN cell. Note that ultimately, this strategy outperforms both the centralized and dis-

[4] Attenuation exponents in a typical urban environment are $\alpha \gtrsim 3.5$.

	Requests' arr. rate $\left(\text{files}\cdot\text{s}^{-1}\cdot\text{km}^{-2}\right)$	File size (Bytes)	Backhaul b/w at WLANs (kbps)
Notation	$\lambda(y)$	$f(y)$	B_k^h
$y \in S_1^w$ or $k \in S_1^w$	$[15,40]$	10	0
$y \in S_n^w$ or $k \in S_n^w$, $n \neq 1$	15	10	500

Table 5.1: Simulation parameters.

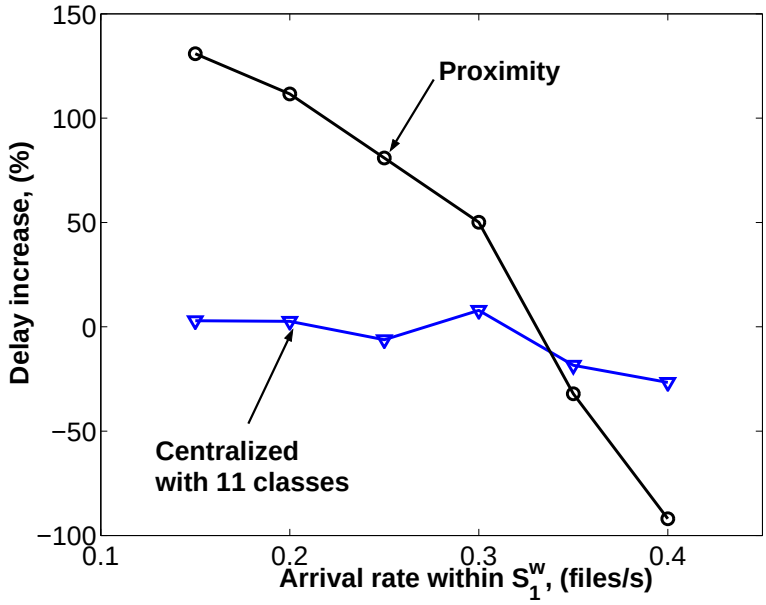


Figure 5.6: Mean delay increase over distributed routing, for centralized and proximity-based routing.

tributed decision-making approaches which operate using the interference-unaware algorithms of Section 5.3.

The key in understanding the seemingly paradoxical outcome of this experiment is as follows. The negative effect on the service capacity of the WAN AP w_1 from interference of surrounding WAN APs is the largest when w_1 approaches high utilization levels due to the increased load on S_1^w. The interference occurs only when WAN APs w_m, $m \neq 1$ are active, i.e. they serve requests that have emerged in S_m^w, $m \neq 1$ and have been routed to the WAN. Thus the decision-making within service zone S_m^w, $m \neq 1$ potentially, crucially effect the performance within S_1^w when w_1 is close to the edge of its servicing capability.

Now, note that the decision-making strategies, obtained in Section 5.3 have been for-

mulated in such a way that they force each WAN service zone to make routing decisions in a greedy fashion, i.e. each WAN AP tries to balance loads within its service zone only. Note that in symmetric load scenarios, the degradation in service capability due to WAN APs activity is similar across service zones, and thus eventually, agents within different WAN service zones will "feel" equal degradation in the WAN service capacity due to interference, which will cause them to make more considerate towards their neighbors routing decisions, i.e. they will be forced to route more requests towards the WLANs.

In our example, however, we have that the WAN AP w_1 is affected much more by the decision-making in surrounding WAN service zones, than these WAN service zones are affected by the decision-making within S_1^w. It follows then that WAN APs w_m, $m \neq 1$ are not in any way stimulated to route more requests to WLAN service zones in order to reduce the interference they cause within S_1^w.

5.4.2 Correction of decision-making to account for other-cell interference

The experiment presented in Section 5.5 suggests that to achieve better performance, the decision-making routing algorithm operating within a particular service zone should factor the effect of its decisions on the performance within the nearby service zones. Another conclusion that one might reach is that it might be beneficial for the system to introduce additional bias causing more requests to be routed to WLANs, under some high and asymmetric load scenarios. How should one design decision-making, in order that system be able to automatically identify such extremal circumstances, and "switch" on and off such biases is the focus of this section.

We return to our general model, in which the WAN service rate at each location depends on the current set of active WAN APs. The queueing dynamics can thus be represented by that of multi-class processor sharing queues, where the service rates at each queue vary over time as governed by the activity state of other queues. Rigorous analysis of such queueing systems is a hard task and the analysis is barely tractable even for two single class queues [73, 74]. Thus, we will adopt the approach introduced in [75] that relies on approximating the performance.

We will first briefly recall the methodology of [75]. The authors consider a multi-cell WAN system with Poisson traffic and processor-sharing service at access points, i.e. their setup is quite similar to the one we have but there is no WLAN layer. The authors start by considering a particular WAN AP, say w_1, in our notation, and assume that the activity pattern of the WAN APs $\{w_m\}_{m\neq 1}$, follows a stochastic process, which is ergodic and independent of the activity of w_1. Based on this preliminary analysis, they construct approximate[5] lower and upper bounds for the mean number of transfers in progress at WAN AP w_1.

The main idea in constructing the upper bound (similar reasoning applies to constructing

[5]The authors don't claim that the bounds hold in general. Their argument is however supported by simulations and the fact that bounds are valid for a particular case of single-class queues.

the lower bound) is to note that the mean number of transfers at WAN AP w_1 is the largest, when the activity pattern of WAN APs w_m, $m \neq 1$ is as in some "worst case scenario". The "worst case scenario" activity pattern can be taken as the one where the nearby WAN APs are active at all times, since clearly, the other-cell interference within S_1^w is the largest in that case and thus the service rates within S_1^w are the lowest. However, the simulations provided in [75] indicate that a bound corresponding to such definition of "worst case scenario" might be quite loose.

In order to improve the bound the authors note that, in fact, one can consider another definition of "worst case" activity pattern. In this new construction, the service rates at WAN APs w_m, $m \neq 1$ are as the ones corresponding to the case where all other WAN APs in the system are active. Denoting $\mathcal{A}_1(t)$ the resulting activity pattern of all WAN APs in the system except for WAN AP w_1, the improved upper bound on the mean number of transfers within S_1^w is obtained under assumption that WAN AP w_1 operates under interference conditions induced by the activity process $\mathcal{A}_1(t)$. The actual upper bound value is then computed using the so-called quasi-stationary regime approximation, in which the queuing dynamics at WAN AP w_1 is much faster than the changes in activity pattern of WAN APs w_m, $m \neq 1$.

At this point we hope that we have demonstrated that the system we are dealing with exhibits enough complexity so that we are motivated to take an approach which relies heavily on approximations and which follows steps that are not rigorously argumented but seem, nevertheless reasonable. The steps we will follow are somewhat similar to the ones described above [75], in terms of the assumptions that we make. However, there are some differences associated with the specifics of this study – by contrast to [75] we are not interested in obtaining closed-form solutions for performance measures. Thus, in our case there will be no need to assume that activity pattern of WAN APs follows any "worst case scenario" process – in the algorithms that we will suggest in a short while, the measurements that depend on the actual realized activity pattern will be incorporated into decision-making.

To derive our new, interference-aware, load-balancing algorithms, we will make the following assumptions.

Assumption 5.1. *For any fixed strategy τ, the stochastic process $\mathcal{A}^\tau(t)$, representing the dynamics of activity of all WAN APs in the system is stationary and ergodic, and has the same distribution as $\mathcal{A}^\tau$.*

Clearly, realistic systems might not obey Assumption 5.1 for all possible control strategies, τ, as the process $\mathcal{A}^\tau(t)$ in general depends on the strategy τ itself. Assumption 5.1 effectively restricts us to work only with strategies that lead to validity of this assumption.

Assumption 5.2. *The system operates in quasi-stationary regime, i.e. the queueing dynamics within the service zone of each WAN AP is much faster than the changes in the set of active WAN APs.*

Using Assumption 5.2 we can readily obtain the mean number of transfers in progress at WAN AP w_m:

$$\mathbb{E}^{\tau}[Q_m^w] = \mathbb{E}_{\mathcal{A}^{\tau}}\left[\frac{\rho_m^w(\tau, \mathcal{A}^{\tau})}{1-\rho_m^w(\tau, \mathcal{A}^{\tau})}\right],$$

where $\rho_m^w(\tau, \mathcal{A}^{\tau})$ is utilization of WAN AP w_m under decision-making strategy τ, when the set of active WAN APs is fixed and is represented by $\mathcal{A}^{\tau}$, and we use $\mathbb{E}_{\mathcal{A}^{\tau}}[\cdot]$ to denote the expectation with respect to the distribution of $\mathcal{A}^{\tau}$. For example, using the notation introduced in Section 5.3.3, we have that τ_1 corresponds to routing vector $\mathbf{p}$, and the utilization $\rho_1^w(\tau, \mathcal{F})$ can be expressed as a function of $\mathbf{p}$ as:

$$\rho_1^w(\mathbf{p}, \mathcal{F}) = \sum_{i=1}^{L} \frac{\gamma_i |\Delta S_i| p_i}{B_{\mathcal{F}}^w(y_i)}$$

where $B_{\mathcal{F}}^w(y)$ denotes the average WAN physical data rate available at location y when the set of active WAN APs is $\mathcal{F}$.

The next assumption is less crucial for our algorithms' derivation and is mostly motivated by the resulting implementation simplifications that it enables.

Assumption 5.3. *The service rates at WAN AP w_m are affected only by the activity pattern of the set $\mathcal{N}_m$ of WAN APs that are immediate neighbors of w_m, i.e., the ones that share service zone boundary with w_m.*

In what follows with some abuse of notation, we will use $\mathcal{A}_m^{\tau}(t) \subset \mathcal{N}_m$ to refer to the set of WAN APs that are immediate neighbors to w_m and, for fixed control τ are active at time t (correspondingly, we will also use $\mathcal{A}_m^{\tau}$ to refer to the random set that has the same stationary distribution as $\mathcal{A}_m^{\tau}(t)$).

We now return to the system objective that is given by equation (5.3). As in Section 5.3, we can still decompose τ into a family of strategies $\{\tau_m\}_{m=1}^M$, so that τ_m operates on the requests emerging solely within S_m^w. The difference from Section 5.3 is that $\{\tau_m\}_{m=1}^M$, which would minimize the system objective over some suitable set of decision-making strategies, can no longer be treated as independent across different m. Thus, in general, the system optimization can not be posed separately for each individual WAN service zone. This leads to certain implementation problems, since even if the solution is possible to obtain it would require a controller that needs to coordinate potentially among all WAN APs in the system. Instead, we will focus on designing strategies $\{\tau_m\}_{m=1}^M$ which are independent, but in which τ_m takes into account some feedback from WAN APs $w_n \in \mathcal{N}_m$.

In order to achieve this goal we define "partial" system objectives $\widetilde{U}_m(\tau)$, corresponding

to each service zone S_m^w, $m = 1, 2, \ldots, M$:

$$\widetilde{U}_m(\tau) = \mathbb{E}^{\tau}[Q_m^w] + \sum_{k \in \mathcal{K}_m} \mathbb{E}^{\tau}[Q_k^h] + C_{IF} \sum_{n \in \mathcal{N}_m} \mathbb{E}^{\tau}[Q_n^w],$$

where C_{IF} is some positive constant. We impose that the decision-making entity within S_m^w strives to minimize $\widetilde{U}_m$ by adjusting decision rules τ_m. Note that, in comparison to the "interference unaware" objectives $U_m(\tau)$ which we used in Section 5.3, the objectives $\widetilde{U}_m(\tau)$ contain additional factor proportional the total queueing backlog of WAN APs neighboring to w_m. In this way we try to enforce that the decision-making within S_m^w is sensitive to the queuing backlog at the neighboring WAN APs.

Note that decision rules τ_m are still potentially adjusted as a reaction to the change in decision rules τ_n, $n \neq m$. So in fact, the dynamics of decisions' adjustments represents a game, in which the interacting entities adjusting decision-making rules are the players, and where each player adjusts its strategy in order to maximize its own utility, represented by the corresponding partial objective. The dynamics might have fixed points (Nash equilibria in game-theoretic terminology), i.e. sets of policies $\{\tau_m^*\}_{m=1}^M$, such that no player desires to adjust it strategy. However proving existence of such equilibria and their stability is a hard task by itself, and will be out of scope of this work. In what follows we will simply experimentally validate that such equilibrium decision-making strategy exists and the long-term decision-making dynamics is eventually sufficiently close to this strategy.

The last preliminary step that we take is making assumptions that allows us to describe the actual changes that occur within the system due to a small variation of a single component of vector τ, say τ_1. Note that for $m \in \mathcal{N}_1$ we can rearrange terms of $\mathbb{E}^{\tau}[Q_m^w]$ to obtain:

$$\mathbb{E}^{\tau}[Q_m^w] \quad = A_1^{\tau} \mathbb{E}_{\mathcal{A}_m^{\tau}}[Q_m^w \mid w_1 \text{ is busy}] + (1 - A_1^{\tau}) \mathbb{E}_{\mathcal{A}_m^{\tau}}[Q_m^w \mid w_1 \text{ is silent}], \tag{5.12}$$

where we denoted via A_1^{τ} the probability that, under control strategy τ, the WAN AP w_1 is active. The conditional expectation in this expression correspond to taking the average, under particular fixed control strategy τ, of quantity Q_m^w with respect to activity pattern of all WAN APs within $\mathcal{N}_m$ except for the WAN AP w_1. Thus in expression (5.12) the effect corresponding to the activity of WAN AP w_1 is factored separately from the effect of the activity of all other WAN APs neighboring to w_m. Thus we need to understand how these two effects are influenced by τ_1 in order to deduce how $\mathbb{E}^{\tau}[Q_m^w]$ changes due to variations in τ_1.

Note first, that variations in τ_1 are likely to affect the amount of traffic routed to WAN AP w_1 and thus do potentially change the activity probability A_1^{τ}. In turn, the change in activity of WAN AP w_1 would also potentially affect the activity of WAN APs neighboring to w_m, since the WAN APs w_n, $n \in \mathcal{N}_1 \cap \mathcal{N}_m$ will "see" a different interference pattern from w_1 within their

service zones and thus would have a changed set of service capacities. The latter would result in changed distribution for $\mathcal{A}_m^{\tau}(t)$.

In this regard, we will make another crucial assumption, under which we will consider the effect on $\mathbb{E}^{\tau}[Q_m^w]$ associated with the change of distribution $\mathcal{A}_m^{\tau}$ on $\mathbb{E}^{\tau}[Q_m^w]$, occurring due to change in τ_1, to be negligible, in comparison of the effect on the same quantity of the change in activity probabilities $A_1(\tau)$ induced by the change in τ_1. Expressing this in mathematical form we have:

Assumption 5.4.

$$\delta_{\tau_1}\mathbb{E}^{\tau}[Q_m^w] \quad = \left(\mathbb{E}_{\mathcal{A}_m^{\tau}}[Q_m^w \mid w_1 \textit{ is busy}] - \mathbb{E}_{\mathcal{A}_m^{\tau}}[Q_m^w \mid w_1 \textit{ is silent}]\right)\delta_{\tau_1}A_1^{\tau} + o(\delta_{\tau_1}A_1^{\tau}),$$

where $\delta_{\tau_1}F$ is a first variation of F due to the variation in τ_1.

We are now in position to derive our interference-aware load balancing rules. Note, that the very construction of the "partial" objectives implies that the decision-making rules will be similar to the ones that we obtained in interference-free case, but will involve some additional "correction" terms which would reflect the performance penalties incurred due to other-cell interference.

5.4.3 Correction term in distributed estimation of the gradient

We will derive the correction to load-balancing algorithm for the case of distributed implementation (it can be done similarly for the centralized version of the algorithm). Under Assumptions 5.1-5.4 we can concentrate on decision-making within a particular service zone, and we choose to focus on S_1^w, as before. Thus we will use the notation of Section 5.3.3, under which the policy τ_1 corresponds to the vector of routing probabilities $\mathbf{p}$.

Using Assumption 5.4 the convexity of the function $\widetilde{U}_1(\mathbf{p})$ with respect to the vector $\mathbf{p}$ could be established, however due to our other assumptions it might be expected to hold only "approximately" in reality. The elements of the gradient of the partial objective can now be expressed as:

$$\widetilde{G}_i \equiv \frac{\partial \widetilde{U}_1}{\partial p_i} = \gamma(y_i)\Delta S_i\left(\widetilde{G}_i^w - G_i^h\right),$$

where G_i^h is the same as in (5.11), and $\widetilde{G}_i^w$ includes a correction term, which involves "cost of interference" caused to neighboring cells:

$$\widetilde{G}_i^w = \mathbb{E}_{\mathcal{A}_1^{\tau}}\left[\frac{1}{B_{\mathcal{A}_1^{\tau}}^w(y_i)(1-\rho_1^w(\mathbf{p},\mathcal{A}_1^{\tau}))^2}\right]$$

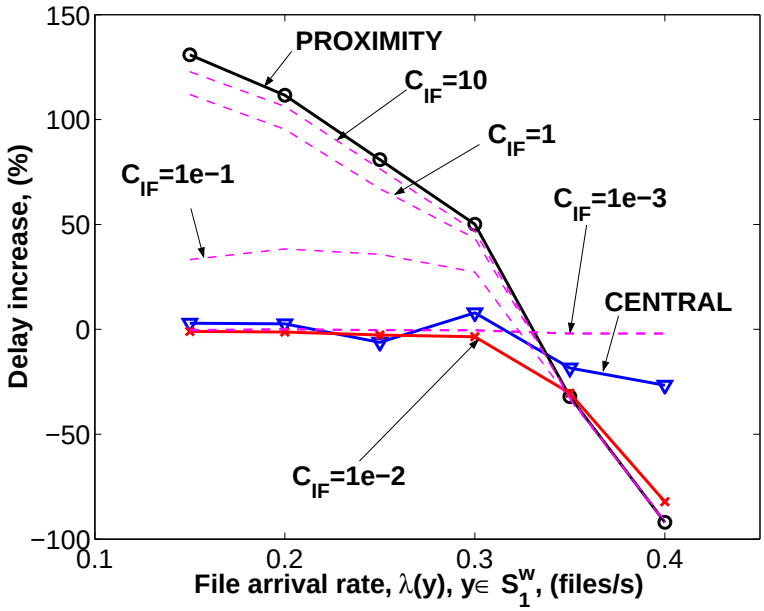


Figure 5.7: Mean delay increase over interference-unaware distributed decision-making for other decision-making strategies, under asymmetric loads (path loss, $\alpha = 3.5$)

$$+C_{IF}\mathbb{E}_{\mathcal{A}_1^\tau}\left[\frac{1}{B^w_{\mathcal{A}_1^\tau}(y_i)}\right]\sum_{m\in\mathcal{N}_1}\left(\mathbb{E}_{\mathcal{A}_m^\tau}[Q^w_m \mid w_1 \text{ is busy}] - \mathbb{E}_{\mathcal{A}_m^\tau}[Q^w_m \mid w_1 \text{ is silent}]\right). \quad (5.13)$$

The interference-aware decision-making algorithms are the same as we described in Section 5.3.3, except that an agent at $y \in \Delta S_i$ selects to join the WAN or WLAN AP via comparing $\widetilde{G}^w_i$ with G^h_i. To enable the computation of $\widetilde{G}^w_i$ the WAN access points have to be able to inform their neighboring WAN APs of the changes in their activity – such information could be signaled over a dedicated wireline or wireless channel. With such coordination between APs, the WAN AP w_m neighboring to, say, w_1 will be able to estimate the averages $\mathbb{E}_{\mathcal{A}_m^\tau}[Q^w_m \mid w_1 \text{ is busy}]$ and $\mathbb{E}_{\mathcal{A}_m^\tau}[Q^w_m \mid w_1 \text{ is silent}]$ by averaging its queue length over times when w_1 is active or silent respectively[6]. These estimates will then be signaled to WAN AP w_1 and will enable computation of $\widetilde{G}^w_i$.

5.5 Performance of interference-aware decision-making

Figure 5.7 shows the performance of the "corrected interference-aware" decision-making in the "paradoxical" setup described in Section 5.4.1. The "partial" objectives used in our formulation of the interference-aware algorithms use the current queue lengths at all neighboring WAN APs

[6]Note that if the adaptation of vector $\mathbf{p}$ is much slower than dynamics of changes in activity of WAN APs, the estimation of the expectations $\mathbb{E}_{\mathcal{A}_m^\tau}[\cdot]$ in expression (5.13) can be done in a straightforward way by maintaining a finite history of measurements for respective quantities.

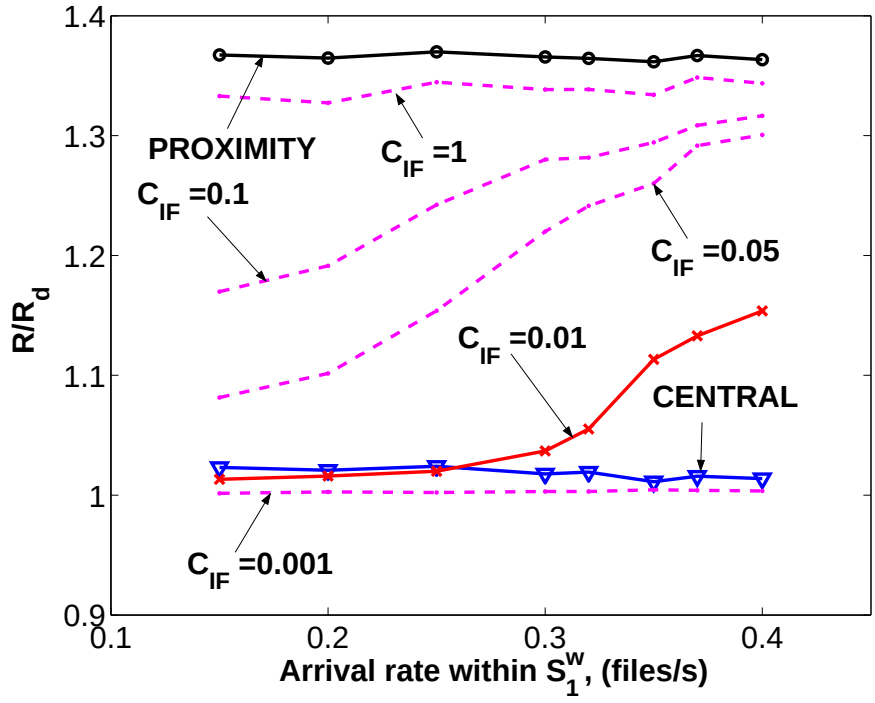


Figure 5.8: Average WAN rates within S_1^w (as multiples of average WAN rate under interference-unaware distributed decision-making) under different values of C_{IF} for increasing load within S_1^w.

to signal their congestion. Including these signals is equivalent to inducing a bias on the system that forces more requests to be routed towards WLANs when the whole system can benefit from it. Note that the value of this bias can be controlled by tuning the constant C_{IF}. We illustrate this tuning in Figure 5.7 by showing performance for different C_{IF}. Note that with properly "tuned" interference-aware decision-making it is possible to achieve both the gains of distributed load-balancing under light loads and proximity-based routing, for heavy asymmetric loads.

Figure 5.8 illustrates the behavior of average WAN rates for different bias constants C_{IF}, when we increase the load within S_1^w. Note that for proximity-based decision-making the average WAN rates are the greatest, since the number of requests routed to WAN APs is the least, and thus the other-cell interference is the least. Note that when the central cell starts to approach the limit of its capacity, the "extra" service rate gained from routing more requests at adjacent WAN service zones to the corresponding hotspots prevents the central cell from the point of breaking down. This is illustrated by the upward moving of the average WAN rates corresponding to the interference-corrected decision-making.

Good performance of interference-aware strategies has been verified also in other experiments. The gains from employing interference-aware policies, however, depend on environ-

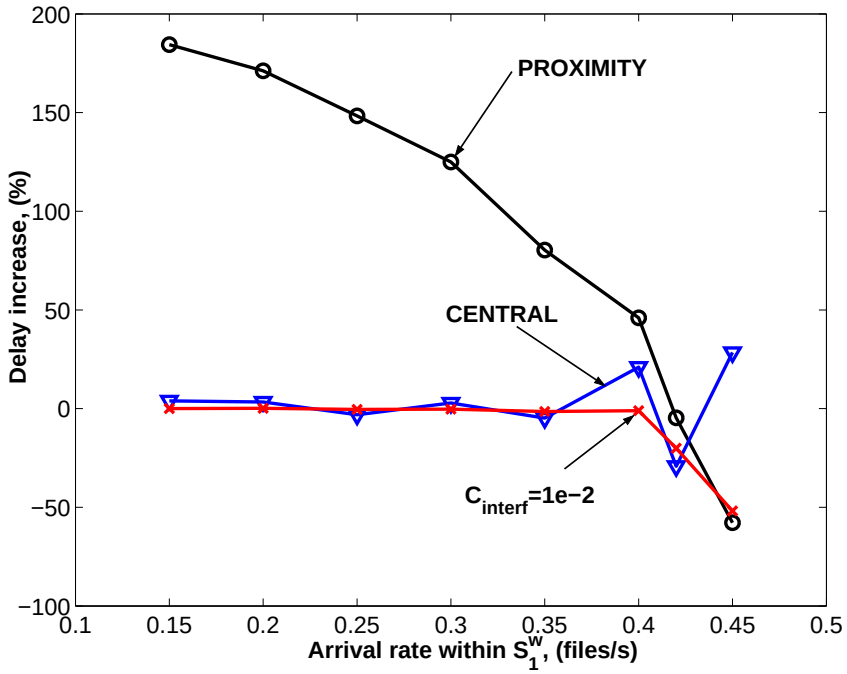


Figure 5.9: Mean delay increase over interference-unaware distributed decision-making for other decision-making strategies, under asymmetric loads (path loss, $\alpha = 3.5$).

mental propagation characteristics, degree of load asymmetry and particular capacities available at various access points in the system. In Figure 5.9 we exhibit the outcome of an experiment that has the same geometric and traffic parameters, as the one described in Section 5.4.1, except that we have set the value of path loss exponent, α to 3.5. Note that the interference-aware decision-making still exhibits gains, which are, however more moderate in comparison to the case with $\alpha = 2$. This outcome is quite expected, since larger attenuation factors lead to diminished effect of the other-cell interference on the capacity of WAN service zones.

At Figure 5.10, we show the geometric setup of another experiment, which has a "larger-scale" heterogeneous network with a number of interfering WAN service zones. The traffic pattern resembles a chessboard: lightly loaded WAN cells are intermittent with heavily loaded WAN cells. Figure 5.11 exhibits the gains that are achieved when employing an interference-aware decision-making vs. a simple distributed decision-making. It is interesting to note that in this particular setup both the proximity-based and interference-unaware centralized decision-making strategies perform quite badly, leading to unstable queueing dynamics – this is why we do not show the performance of these strategies at Figure 5.11.

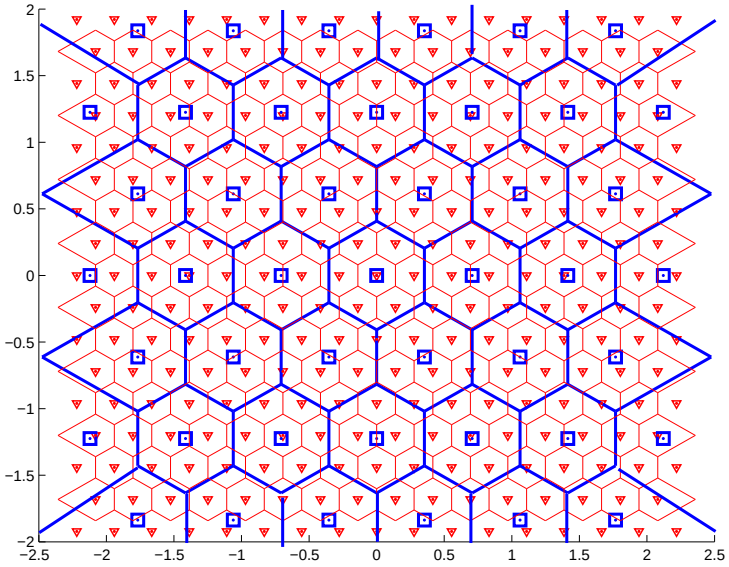


Figure 5.10: Geometric setup for an experiment.

5.6 Conclusion

In this chapter we presented our results on achieving load-balancing among heterogeneous wireless systems that include a combination of WAN and a set of WLANs. Under the assumption that the access points serve the incoming download requests in a processor-sharing fashion, we formulate and evaluate centralized and distributed decision-making routing schemes that enable significant performance gains. The major contribution of this work is explicit incorporation of other-cell interference as a factor affecting load-balancing decisions and exhibiting scenarios in which interference-aware algorithms achieve significant performance gains.

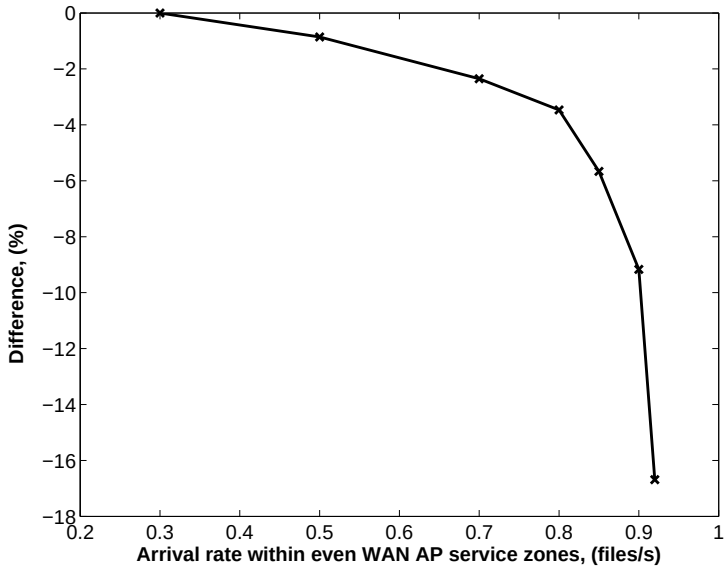


Figure 5.11: Percent delay difference for interference-aware decision-making and interference-unaware decision-making

Part II

Hybrid Networks

Chapter 6

Capacity of hybrid wireless ad hoc networks with infrastructure support

6.1 Introduction

In this chapter we investigate the per user throughput that can be achieved in a hybrid ad hoc wireless network. The network consists of ad hoc nodes, that can relay information among each other via wireless links, and of infrastructure nodes, that can communicate with ad hoc nodes in a wireless manner but are also interconnected via independent high capacity wired or wireless links. We shall refer to the latter as infrastructure nodes or base stations interchangeably. There has been extensive interest in studying purely ad hoc networks, for applications involving networking of military, emergency services and, more recently, to enable the inexpensive deployment of large numbers of networked sensors in the field [76, 77, 78, 79]. Since wireless units are typically energy constrained and ad hoc networks may have limited communication capacity, the addition of infrastructure or base station nodes is a natural approach to reducing the energy and traffic burden on ad hoc nodes while possibly increasing the system throughput. For example, one can envisage a hybrid ad hoc wireless network as a means to enable sharing of information between possibly mobile sensor nodes or gathering of sensed information towards query points on a wireline network. In this case infrastructure nodes are leveraged by ad hoc nodes to reduce their energy burden and increase capacity. Alternatively one can view hybrid ad hoc network model as a means to extend the communication coverage of wireless cellular infrastructure. In this case base stations would leverage spatially distributed ad hoc nodes that are willing to relay information to increase coverage and, possibly, capacity. Note that by contrast with cellular systems in a hybrid ad hoc network not all traffic needs to be mediated through a base station, i.e., wireless nodes that wish to communicate with each other might do so directly. Thus there are two "types" of traffic in such networks: that which is eventually mediated through the infrastructure nodes and that which is relayed in a purely ad hoc manner.

In this chapter we will study the per user throughput scaling laws as the numbers of ad hoc and infrastructure nodes in a hybrid ad hoc network grow.

6.1.1 Related work

Let us briefly consider what is known for purely ad hoc networks [80, 19, 81, 82, 83, 84, 14, 21, 85, 86]. The first key result is that under a reasonable interference model the aggregate transport capacity of an *arbitrary network* scales as $\Theta(W\sqrt{An})$ bits-meter/sec in the number of nodes n each placed in a region with area A where each has capacity W bits/sec [19]. An arbitrary network is one in which the placement of ad hoc nodes and traffic loads can be selected so as to maximize the capacity. The insight offered in [19] is that system throughput is maximized when one minimizes the transmission power of each node resulting in a quadratic reduction in the interference region of a transmission[1]. Minimizing the interference region permits one to *schedule* as many non-interfering concurrent transmissions as possible resulting in maximized transport capacity. The information theoretic results in [83] suggest that this same basic characteristic capacity scaling will continue to apply under more general communication models with more powerful interference cancellation techniques. In addition to the arbitrary network model [19] studies a *random network* model where nodes are placed at random within a given region, and choose to communicate with random destinations. In this context they study the asymptotic scaling for the minimum *per user throughput*, and show it to be $\Theta(W/\sqrt{n\log n})$ bits/sec. This result provides a better sense of the performance seen by individual nodes under a random traffic pattern exhibiting no spatial locality. Several follow-on papers have shown the effect of traffic locality [81], the impact of directional antennas on capacity [21], and how exploiting mobility might increase the per user throughput with varying penalties in delay [84, 85, 86, 87].

More recently [14] have studied the *aggregate throughput capacity* for a hybrid network model supporting uniformly distributed traffic loads. The model assumes that the base stations are placed on a regular grid, but the ad hoc nodes are randomly distributed within the area of the network. Since the goal of [14] is to bound the aggregate throughput, in their analysis certain users' traffic is shut off, to favor others, e.g., those with "local" traffic. Several additional interesting questions follow from this work. First their analysis is based on partitioning bandwidth statically and in a spatially homogenous manner for purposes of supporting ad hoc and infrastructure communications. Thus one might ask whether a spatially dependent partitioning might not result in a higher aggregate or per user throughput. Indeed, since infrastructure nodes are wired together, they will be very desirable as a shortcut for traffic that needs to traverse a long distance. Thus one would expect infrastructure nodes to become traffic hotspots. One might then question whether allowing a larger portion of bandwidth for communication with

[1] Loosely, the interference region of a transmission is the area around it wherein it would interfere with other nodes' receptions.

infrastructure nodes, in regions close to such nodes while increasing the bandwidth allocated for ad hoc communications far away from these nodes would not improve the system capacity. In [14] the capacities of two natural routing strategies between ad hoc and infrastructure nodes are studied. Thus a second question arises as to whether a better routing strategy, e.g., one that avoids hot spots, for traffic to be carried in the ad hoc mode, might not be able to achieve a better asymptotic capacity scaling.

In contrast to [14] the work of [88] considers a hybrid network model where the positions of both base stations and ad hoc nodes are randomly selected. The authors determine the asymptotic scaling of capacity for such a network when the number of base stations scales *linearly* with the number of ad hoc nodes. No other scaling regimes are considered.

6.1.2 Contributions and organization

In this chapter we resolve some of the issues that are not treated in [14, 88] by proving an upper bound scaling for the *per user throughput* in a random hybrid ad hoc network that is independent of the routing strategy. This bound is shown for networks in which infrastructure nodes are placed in any deterministic manner, including the placement on the regular grid described in [14], but ad hoc nodes are placed at random within a given region. In contrast to [14] where the range of all nodes is the same for all nodes, we allow each infrastructure node to *adjust the range* for each transmission. Therefore, the model is quite general in that it allows one to operate infrastructure nodes in an optimal manner. In contrast to [88], we identify three scaling regimes for the growth of the number of infrastructure nodes, m with the number of ad hoc nodes n. In each of these regimes we demonstrate that the upper bound is tight by specifying a particular placement for infrastructure nodes accompanied with a scheduling and routing strategy.

Specifically we show that whenever $m \lesssim \sqrt{n/\log n}$ the per user throughput is of order $W/\sqrt{n\log n}$. This is the same as that shown for pure ad hoc networks by [19] whence in this regime infrastructure nodes do not enhance the capacity scaling for the hybrid network. Whenever $\sqrt{n/\log n} \lesssim m \lesssim n/\log n$ the order of the per user throughput is Wm/n. Thus in this case the additional aggregate bandwidth Wm brought in by infrastructure nodes appears to be shared by ad hoc nodes. In fact we demonstrate that in this regime, this scaling is achieved by letting ad hoc nodes communicate directly via infrastructure nodes, i.e., as would be the case in a cellular system. Thus more advanced routing and scheduling schemes, as, e.g., proposed in [14], might achieve better leading factors, but would not change the basic asymptotic capacity scaling for such systems. Finally, whenever $m \gtrsim n/\log n$ the achievable per user throughput is of order $W/\log n$. This implies that further investments in the infrastructure do not lead to improvement in the scaling of per user throughput. In some sense this is a result of our random uniform traffic model we are using rather than an intrinsic property of the network itself. However, it is

interesting to note that due to the limited local communication capacity of ad hoc and hybrid networks fluctuations in traffic will lead to significant performance penalties. By contrast, a wireline network with shared but large links can often absorb such fluctuations through statistical multiplexing.

The organization of the chapter is as follows. In Section 6.2 we describe our network and traffic models, and define the feasible throughput for a random hybrid network. In Section 6.3 we show an upper bound on the scaling of per user throughput which is independent of routing. Then, in Section 6.4, we demonstrate that upper bounds are in fact tight. Section 6.5 contains some concluding remarks.

6.2 Hybrid Ad hoc Networks: Model and Notation

6.2.1 Model for a random hybrid network

We consider a network with n ad hoc nodes, that are randomly, i.e., uniformly, placed at locations $X_1, X_2, \ldots, X_n$ within a disc of unit area D. These nodes are capable of transmitting and receiving W bits/sec via a wireless channel. In addition, there are $m(n)$ *arbitrarily placed*[2] infrastructure nodes (or base stations), which are interconnected through wired or alternative wireless links which can support as much traffic as necessary. For simplicity, we shall assume base stations have a capacity of W bits/sec to communicate with ad hoc nodes. In contrast to ad hoc nodes, the base stations do not generate any traffic themselves, i.e., they serve purely as infrastructure which relays traffic on behalf of nodes in the ad hoc network. We will assume that the channel capacity can be split into an arbitrary number l of orthogonal (non-interfering) subchannels each with W_j bits/sec such that $\sum_{j=1}^{l} W_j = W$. In our model each ad hoc or infrastructure node can then send simultaneously using any number of the subchannels, to any number of receivers. However we assume, that an ad hoc node can receive from only a single sender and cannot send and receive in the same instant. We do not impose the latter constraints on the infrastructure nodes, i.e., they are free to send and receive concurrently on orthogonal channels. In the sequel we will consider a sequence of such networks where $n \to \infty$ and $m(n)$ grows according to various scalings – in most cases we write $m(n)$ simply as m. We will limit the cases of interest to that where $m(n) \leq n$ i.e., there are fewer infrastructure nodes than ad hoc nodes.

[2] i.e., placed in any pre-specified, but deterministic way which does not depend on the realization of ad hoc nodes.

Uniform traffic model

We assume that each ad hoc node, say the i-th one at location X_i selects a destination for its traffic as follows. It will choose a random location on the disk S_i and will then choose the node among $X_1, \ldots, X_n$ which is closest to S_i. We shall denote the location of the i-th node's receiver by Y_i, and let L_i denote the length of the segment between X_i and Y_i. For this traffic model it can be shown that L_i, $i = 1, \ldots, n$ will be independent and identically distributed [19]. This traffic model exhibits no locality.

Interference Model

We will adopt the Protocol Model for interference among nodes sharing a wireless channel, see second version in [20]. In particular, consider an ad hoc or infrastructure node located at X_i transmitting towards another (ad hoc or infrastructure) node located at X_j along the subchannel s and using the range r_i for this transmission. The transmission will be successful at X_j if:

$$|X_i - X_j| \leq r_i\,, \text{ and } |X_k - X_j| \geq (1+\Delta) r_k\,, \tag{6.1}$$

where X_k is the location of any other node which is concurrently transmitting over the subchannel s, and $\Delta > 0$ is some parameter. Note, that the first condition in (6.1) requires that the receiver is within the range of the sender, and the second – ensures that the interference caused by any other concurrent transmission in the system is limited at the receiver X_j. We will further restrict the Protocol Model to the case where ad hoc nodes employ a common range r but allow infrastructure nodes that might not be power constrained to adjust their transmission range arbitrarily.

Per user throughput capacity

Following [14] we will extend the definition of [19] for the feasible per user throughput to the case of a random hybrid network with arbitrarily placed infrastructure nodes.

Definition 6.1. *A throughput of $\lambda(n,m)$ per node/user is feasible if there is a placement rule for the base stations, and a spatial and temporal scheme for scheduling transmissions allowing buffering at intermediate nodes (if necessary), such that each node can send $\lambda(n,m)$ bits/sec on average to its chosen destination. That is, there is $T < \infty$, such that in every interval $[(i-1)T, iT]$ every node can send $\lambda(n,m)T$ bits to its corresponding destination node.*

Now, as in [19], we can define the asymptotic scaling for the per user throughput as follows:

Definition 6.2. *The per user throughput $\Lambda(n,m)$ of a random hybrid network is of order*[3] $\Theta(\lambda(n,m))$ *bits/sec if there exist deterministic constants $c > 0$ and $c' < \infty$ such that*

$$\lim_{n\to\infty} \mathbb{P}(\Lambda(n,m) = c\lambda(n,m) \textit{ is feasible}) = 1$$

$$\liminf_{n\to\infty} \mathbb{P}(\Lambda(n,m) = c'\lambda(n,m) \textit{ is feasible}) < 1.$$

These two conditions can be interpreted as asymptotic lower and upper bounds over random realizations for the locations of ad hoc nodes and destinations of the traffic.

6.3 Upper Bound on Per User Throughput

Our upper bound on per user throughput for random hybrid networks will draw on various results in previous work. So we shall start by providing a summary of several key results we will use in this chapter.

6.3.1 Background results

It can be shown that the Protocol model for interference (6.1) requires that whenever two simultaneous transmissions occur on the same subchannel and are successfully recovered, the discs of certain radii centered in the receivers must be disjoint [19]. Thus each successful transmission will necessarily occupy a portion of a total area of the disc placing a constraint on the number of successful receptions that can occur on a given subchannel. The first result below is an adaptation of Lemma 5.4 in [19], and provides a bound on the number of simultaneous successful transmissions that can occur in a *pure ad hoc* network.

Lemma 6.1. *Consider a network with n ad hoc nodes arbitrarily placed within the disc of unit area and let all the nodes use the a common transmission range r for their transmissions. Under the Protocol Model (6.1) the number of successful simultaneous transmissions n_{sim} that can occur on a given subchannel is upper bounded by:*

$$n_{sim} \leq \frac{c_1}{\Delta^2 r^2},$$

where $c_1 > 0$ is some constant, independent of Δ and n.

Lemma 6.1 also shows that the number of simultaneous transmissions in a pure ad hoc network will be maximized, whenever the common transmission range is made as small as

[3] We will use the asymptotic notation as discussed in [89], i.e., $O(g(n))$, $\Omega(g(n))$, $\Theta(g(n))$, and $o(g(n))$ and $\omega(g(n))$.

possible. There is, however, a limit on how small this range can be while still keeping network connectivity. This limit is identified by the following theorem, proven in [90].

Theorem 6.1. *Let n nodes be randomly placed on the disc of unit area according to a uniform distribution. Assume that two nodes are connected if the distance between them is smaller than the "connectivity range" $\rho(n)$, where:*

$$\pi\rho^2(n) \triangleq \frac{\log n + c(n)}{n}. \tag{6.2}$$

Then, all nodes are connected in a single cluster with probability 1 as $n \to \infty$ if and only if $c(n) \to \infty$.

The next result, which we prove in Appendix 6.6, establishes a lower bound on the minimum range which would allow connectivity in a random hybrid network where the infrastructure nodes are placed arbitrarily.

Proposition 6.1. *Let n nodes be randomly placed on the disc of unit area according to a uniform distribution and let m additional nodes be placed arbitrarily. Let $r(n,m)$ be a "connectivity range", i.e. the distance within which two nodes are assumed to be connected. Assume that $r(n,m)$ is chosen so that all $n+m$ nodes form a single connected cluster with probability 1 as $n \to \infty$. Then, $r(n,m)$ satisfies:*

$$r(n,m) = \begin{cases} \Omega\left(\sqrt{\frac{\log n}{n}}\right), & \text{if } m = o\left(\frac{n}{\log n}\right), \\ \Omega\left(\frac{1}{\sqrt{m}}\right), & \text{if } m = \Omega\left(\frac{n}{\log n}\right). \end{cases}$$

Note that Proposition 6.1 establishes a lower bound on the required connectivity range under different joint scalings for the number of ad hoc and infrastructure nodes.

6.3.2 Derivation of the upper bound

We will define three scaling regimes for the number of infrastructure and ad hoc nodes:

$$\begin{aligned} &\text{(i)}: \quad m(n) = O\left(\sqrt{\tfrac{n}{\log n}}\right), \\ &\text{(ii)}: \quad m(n) = \omega\left(\sqrt{\tfrac{n}{\log n}}\right) \text{ and } m = O\left(\tfrac{n}{\log n}\right), \\ &\text{(iii)}: \quad m(n) = \omega\left(\tfrac{n}{\log n}\right). \end{aligned} \tag{6.3}$$

Note that Regime (i) corresponds to the case where $m \lesssim \sqrt{\frac{n}{\log n}}$, Regime (ii) – to the case where $\sqrt{\frac{n}{\log n}} \ll m \lesssim \frac{n}{\log n}$ and Regime (iii) – to the case where $m \gg \frac{n}{\log n}$. Our upper bounds on the

asymptotic scaling for the per user throughput in these three regimes are formally stated in the following theorem.

Theorem 6.2. *The throughput per user, $\Lambda(n,m)$, in a random hybrid network under the protocol model is such that for some $c' > 0$ independent of n and m we have:*

$$\lim_{n\to\infty} \mathbb{P}(\Lambda(n,m) = c'\lambda(n,m) \textit{ is feasible}) = 0,$$

where

$$\lambda(n,m) = \begin{cases} \frac{W}{\sqrt{n\log n}} & \textit{in Regime (i)}, \\ \frac{Wm}{n} & \textit{in Regime (ii)}, \\ \frac{W}{\log n} & \textit{in Regime (iii)}. \end{cases}$$

We prove the theorem via two propositions as follows.

Proposition 6.2. *The throughput per user, $\Lambda(n,m)$ for a random hybrid network under the Protocol model satisfies*

$$\Lambda(n,m) = \begin{cases} O\left(\frac{Wm}{n}\right) & \textit{if } m = \Omega\left(\sqrt{\frac{n}{\log n}}\right) \\ O\left(\frac{W}{\sqrt{n\log n}}\right) & \textit{if } m = O\left(\sqrt{\frac{n}{\log n}}\right) \end{cases}. \tag{6.4}$$

Proof. For simplicity we assume transmissions are slotted, with slots of length τ secs. By definition a throughput λ is feasible if over a large period of time, say $[0,T]$, each node would be able to send λT bits over T/τ transmission slots. Note that each node may send some (or all) of its λT bits using base station infrastructure and must send the remainder in ad hoc mode via multi-hop relaying. The idea of our proof is to bound the ad hoc traffic burden on the network that n nodes would produce.

We will allow nodes to send over l non-interfering sub-channels each with bandwidth W_j such that $\sum_{j=1}^{l} W_j = W$. Since a hybrid network consists of both nodes and base stations with the same capability sharing these subchannels, the maximum bits/sec on the up-link to a particular base station is at most W bits/sec. Since there are m such base stations, the total up-link flow in a given slot is at most $Wm\tau$ bits.

Suppose that $\kappa_i^u(t)Wm\tau$ bits transmitted by node i are relayed on up-links to base stations during time slot t. Let $\bar{\kappa}_i^u$ denote the time average of the fractional up-link base station burden

$\kappa_i^u(t)$ associated with node i, i.e.,

$$\bar{\kappa}_i^u \equiv \frac{1}{T} \sum_{t=1}^{\lceil T/\tau \rceil} \kappa_i^u(t)\tau.$$

This quantity must satisfy the following constraints

$$\sum_{i=1}^{n} \bar{\kappa}_i^u \leq 1/2 \text{ and } \bar{\kappa}_i^u W m \leq \lambda. \tag{6.5}$$

The first constraint is due to flow conservation, i.e., the aggregate up-link burden on base stations can not exceed 1/2 since these same bits will need to be sent by base stations on down-links. The second inequality must hold true for all nodes i since each node's feasible throughput will at least exceed its time average up-link burden on the base stations.

The total number of bits N_i transmitted in 'ad hoc' manner from node i to its destination within the time interval $[0,T]$ is thus at least:

$$N_i \geq \lambda T - \sum_{t=1}^{\lceil T/\tau \rceil} \kappa_i^u(t) W m \tau. \tag{6.6}$$

Note by bits sent in 'ad hoc' manner we mean those which are relayed from the source to the destination without the the help of a base station. Let $b(i)$ be an integer indexing the $b(i)^{th}$ bit sent in ad hoc manner by node i and let $h(b(i))$ denote the total number of hops this bit will take to reach its destination. By summing over all such bits that are sent over sub-channel l during time slot t, we obtain:

$$\sum_{i=1}^{n} \sum_{b(i)=1}^{N_i} \sum_{h=1}^{h(b(i))} \mathbf{1}\,[b(i) \text{ performed } h\text{'th hop over subch.} l \text{ at time slot } t] \leq W_l \tau n_{sim}\,,$$

where we denote by n_{sim} the number of simultaneous transmissions that can occur in a hybrid network with the base stations not participating in the relaying. Note that n_{sim} in this case is the same as the number of simultaneous transmissions for a *purely ad hoc network* consisting of nodes randomly and uniformly placed within the disc D. Using Lemma 6.1 and summing over all the sub-channels and time slots one obtains:

$$\frac{1}{T} \sum_{i=1}^{n} \sum_{b(i)=1}^{N_i} h(b(i)) \leq \frac{c_1 W}{\Delta^2 r^2}\,, \tag{6.7}$$

where r is the common range for all ad hoc nodes' transmissions. We define now the following

quantity:

$$l^r(b(i)) = h(b(i))r,$$

which corresponds to an estimate for the length of the path traversed by bit $b(i)$ to reach its destination, i.e., number of hops times radius of each ad hoc transmission. We also define the following per node averages:

$$\bar{l}_i^r = \begin{cases} \dfrac{1}{N_i} \displaystyle\sum_{b(i)=1}^{N_i} l^r(b(i)) & \text{if } N_i > 0 \\ 0 & \text{if } N_i = 0 \end{cases},$$

i.e., the average distance (measured along the path) traversed by bits carried in ad hoc manner on behalf of node i during the time interval $[0,T]$. Thus, from (6.7) we have:

$$\frac{1}{T}\sum_{i=1}^{n} N_i \bar{l}_i^r \leq \frac{c_1 W}{\Delta^2 r}.$$

Note, furthermore, that if $N_i > 0$ then $\bar{l}_i^r$ is lower bounded by l_i, that is the distance, measured along the straight line from the node i to its intended receiver. So, it follows that :

$$\frac{1}{T}\sum_{i=1}^{n} N_i l_i \leq \frac{c_1 W}{\pi\Delta^2 r^2}.$$

Next, using our lower bound (6.6) for N_i, we obtain:

$$\sum_{i=1}^{n}\left[\lambda - \frac{Wm}{T}\sum_{t=1}^{\lceil T/\tau\rceil} \kappa_i^u(t)\tau\right] l_i \leq \frac{c_1 W}{\Delta^2 r}.$$

so,

$$\lambda \leq \frac{Wm\sum_{i=1}^{n} \bar{\kappa}_i^u l_i + \frac{c_1 W}{\Delta^2 r}}{\sum_{i=1}^{n} l_i}. \tag{6.8}$$

By the Strong Law of Large Numbers we have that:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} l_i = \mathbb{E}[L_1], \text{ almost surely},$$

where we took into account the fact that the sequence $\{l_i\}_{i=1}^n$ are realizations of a sequence of i.i.d. random variables which share the distribution of L_1. Since $\mathbb{E}[L_1] = \Theta(1)$, we have that $\sum_{i=1}^n l_i = \Theta(n)$, which combined with (6.8) and the constraints in (6.5) yields the following

upper bound for the feasible rate:

$$\max_{\bar{\kappa}_i^u,(i=1,\ldots,n)} \left[\frac{Wm}{n} \sum_{i=1}^{n} \bar{\kappa}_i^u l_i + \frac{c_1 W}{\Delta^2 nr} \right],$$

$$\text{under constraints: } \sum_{i=1}^{n} \bar{\kappa}_i^u \leq 1/2\,,\ \bar{\kappa}_i^u Wm \leq \lambda\,. \tag{6.9}$$

This optimization corresponds to a Knapsack problem with linear constraints, whence it is optimal to give maximal values $\bar{\kappa}_i^u$'s to nodes i associated with the largest weights l_i. Since $\bar{\kappa}_i^u \leq \frac{\lambda}{Wm}$ and $\sum_{i=1}^{n} \bar{\kappa}_i^u \leq 1/2$ it follows that only the $\lceil \frac{Wm}{2\lambda} \rceil$ largest l_i's out of n can be accommodated – we have relaxed the constraints by taking the ceiling function. Let $\pi(i)$ denote a permutation of the node indices such that $l_{\pi(1)} \geq l_{\pi(2)} \geq, \ldots, \geq l_{\pi(n)}$. Then the upper bound will be maximized when $\bar{\kappa}_i^u$ are set as follows:

$$\bar{\kappa}_{\pi(i)}^u = \begin{cases} \frac{\lambda}{Wm} & \text{for } i = 1, \ldots, \lceil \frac{Wm}{2\lambda} \rceil \\ 0 & \text{otherwise} \end{cases}.$$

Noting that for each i, $l_i = \Theta(1)$, the upper bound for the per user throughput reads:

$$\lambda \leq c_2 \frac{Wm}{n} + c_3 \frac{W}{\Delta^2 nr} \tag{6.10}$$

for two constants $c_2, c_3 > 0$. It is clear then that r must be chosen as small as possible, while ensuring that the the n nodes and m base stations will still be asymptotically connected. Incorporating the lower bounds for r identified by Proposition 6.1 into inequality (6.10) we obtain the asymptotic upper bounds on per user throughput stated in the proposition. □

The next proposition establishes an upper bound on the per user throughput in a random hybrid network irrespective of the number of infrastructure nodes in the network.

Proposition 6.3. *The per user throughput, $\Lambda(n,m)$, for a random hybrid network under the protocol model satisfies, for any m:*

$$\Lambda(n,m) = O\left(\frac{W}{\log n} \right). \tag{6.11}$$

Proof. We will show that with some positive probability, that does not diminish as $n \to \infty$, there exists at least one node which is selected as an intended receiver by $\Theta(\log(n))$ senders. The result will then follow by noting that the receiving capacity of each node is limited by W bits/sec.

We first note that, by Theorem 6.1, the geometric graph $\mathcal{G}^n$ of n nodes placed randomly

within a unit area disc is asymptotically disconnected, if the connectivity radius $\tilde{r}(n)$ is chosen as to satisfy:

$$\pi\tilde{r}^2(n) = \frac{\log n + c}{n},$$

where $c > 0$. Furthermore, by the proof of Theorem 2.1 in [90], the probability $P(n)$ of having at least one isolated node in $\mathcal{G}^n$ is lower bounded by some positive constant for all sufficiently large n. Thus, for all sufficiently large n, with some positive probability, there exists an ad hoc node $I^n \in \mathcal{G}^n$, which has no ad hoc neighbors within a distance $\tilde{r}(n)$.

Second, we show that with probability that does not diminish as $n \to \infty$, the node I^n is chosen to be an intended receiver for as many as $C\log n$ senders, (for some $C > 0$ independent of n). Recall that for each node i, it's intended receiver Y_i is the closest node to a randomly a uniformly distributed location S_i on the disc. The node I^n will for sure be chosen as intended receiver for the i-th ad hoc node if S_i falls within the distance $\tilde{r}(n)/2$ from I^n. The ensemble average of nodes, $\mathbb{E}N(I^n)$, having I^n as their intended receiver is thus:

$$\mathbb{E}N(I^n) \geq \frac{n\pi\tilde{r}^2}{4} = \frac{\log(n)}{4}. \tag{6.12}$$

Next, using the Chernoff bound we show that the deviations from this average are negligible. Indeed if $N(I^n)$ is the actual number of nodes having I^n as their receiver we have that for any $\delta > 0$:

$$\mathbb{P}\Big(N(I^n) < (1-\delta)\mathbb{E}N(I^n)\Big) \leq \exp\left(-\frac{\delta^2\mathbb{E}N(I^n)}{2}\right) \leq \frac{1}{n^{\delta^2/8}}$$

where we have used the inequality (6.12). Thus we have that the actual number of nodes sending to the node I^n with probability tending to 1 as $n \to \infty$ is lower bounded by:

$$N(I^n) \geq C\log(n), \tag{6.13}$$

with some $C > 0$ independent of m and n.

Finally, we note that the receiving throughput capacity of each node is bounded by W bits/sec whether it receives from a base station or from ad hoc nodes. Thus, since each node sends λ bits/sec, the throughput must satisfy the following inequality:

$$\lambda N(I^n) \leq W.$$

Combined with (6.13), the previous inequality yields:

$$\lambda \leq \frac{W}{C\log n},$$

and the statement of the proposition follows. □

Theorem 6.2 combines the various scaling regimes for m and n considered in Propositions 6.2 and 6.3, noting which gives the tightest upper bound.

6.4 Lower bounds on throughput capacity

In this section we will show the upper bounds obtained in Theorem 6.2 are tight in the sense of Definition 6.1. We will consider separately each of the three scaling regimes defined by (6.3).

6.4.1 Regime (i)

The following corollary is an immediate consequence of Proposition 6.2.

Corrollary 6.1. *If $m = O\left(\sqrt{\frac{\log n}{n}}\right)$, the order given in Theorem 6.2 is feasible, i.e., the throughput of a random hybrid network is of order $\Theta\left(\frac{W}{\sqrt{n\log n}}\right)$.*

Proof. The per user throughput of a random ad hoc network is of order $\Theta\left(\frac{W}{\sqrt{n\log n}}\right)$ in [19]. This matches the upper bound in Theorem 6.2 for a random hybrid network whenever $m = O\left(\sqrt{\frac{n}{\log n}}\right)$. Hence this per user throughput is achievable, and can be achieved through ad hoc communications alone, i.e. there is no need to deploy the infrastructure nodes. □

6.4.2 Regime (ii).

For the regime where $\sqrt{\frac{n}{\log n}} \ll m \lesssim \frac{n}{\log n}$ we shall describe a placement of base stations and a routing and scheduling strategy demonstrating that the order of the throughput specified by Theorem 6.2 is feasible. Below we will construct a Voronoi tessellation of the disc D which meets certain regularity properties. The tessellation will be associated with the placement of the infrastructure nodes and be used to partition the disc into regions of operation for each base station – we refer to it as the "infrastructure" tessellation. Subsequently we will show that there exists a scheduling strategy where each node needs only to communicate *exclusively* with its closest base station and can realize its desired asymptotic throughput.

Placement of base stations and the "infrastructure" tessellation

We shall split the plane into a hexagonal tessellation $\mathcal{H}_m$ with the side of a hexagon equal to $\frac{1}{\sqrt{m}}$. The following fact is proven in appendix.

Fact 6.1. *The number m_0 of the hexagons in $\mathcal{H}_m$, which are fully contained in the unit area disc D, satisfies $m_0 < m$ and $m_0 = \Theta(m)$.*

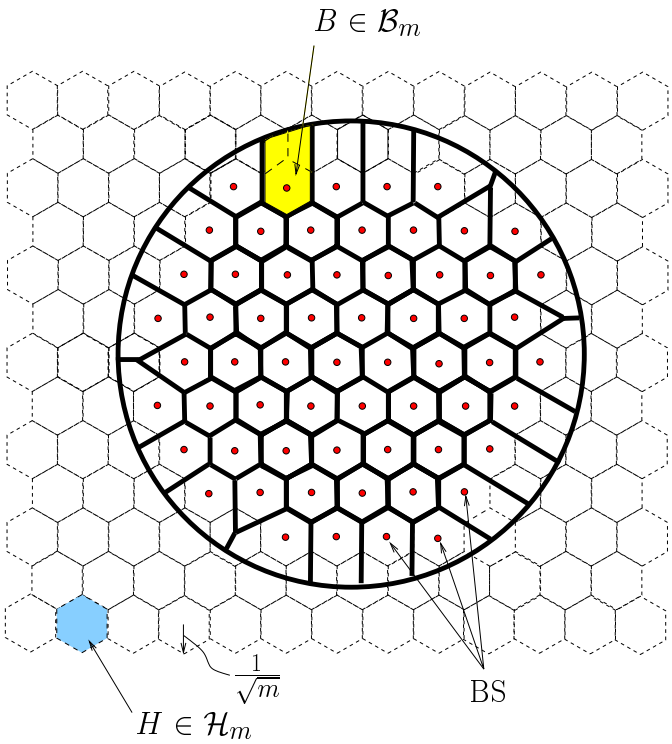


Figure 6.1: The tessellation $\mathcal{B}_m$ induced by the placement of the base stations on the disc.

Suppose we place m_0 out of m base stations into the centers of the hexagons that are fully contained inside the disc D and let $\mathcal{B}_m$ denote the Voronoi tessellation of D induced by these base stations. The remaining $m - m_0$ base stations are left out of our construction (equivalently, those base stations can be arbitrary placed anywhere and shut down). As shown on Figure 6.1, the cells of the tessellation $\mathcal{B}_m$ coincide with the hexagons of $\mathcal{H}_m$, except for those close to the boundary. Based on this construction one can show the following regularity property, see the appendix.

Lemma 6.2. *For a sufficiently large m, each cell B of the tessellation $\mathcal{B}_m$ contains a disc of radius $\sqrt{\frac{3}{2m}}$ and is contained in a disc of radius $\frac{3}{\sqrt{m}}$.*

Routing

Suppose each node directly sends to its closest base station via a wireless channel for routing its packets on the up-link or receiving packets on the down-link. We assign equal ranges

$$r_{i.s.}(m) = \frac{6}{\sqrt{m}}, \tag{6.14}$$

for all nodes transmitting on the up-link and let the base stations use the same range for down-link communications. The following is a simple consequence of Lemma 6.2:

Fact 6.2. *The range $r_{i.s.}(m)$ is sufficient to allow each node to communicate with its closest base station directly, i.e., in a single hop.*

Interference

We now consider interference among nodes transmitting simultaneously. Analogously to [19], we call the "infrastructure" cells $B, B' \in \mathcal{B}_m$ interfering neighbors if there are two points $S \in B$ and $S' \in B'$, such that the distance along the straight line between them, $d(S,S')$ satisfies: :

$$d(S,S') \leq (2+\Delta)r(m)\,.$$

The following lemma is an adaptation of Lemma 4.3 in [19] and is proven in the appendix.

Lemma 6.3. *Let $n(B)$ denote the number of interfering cells for a cell $B \in \mathcal{B}_m$ then,*

$$n(B) \leq C_{\mathcal{B}}\,. \tag{6.15}$$

where $C_{\mathcal{B}}$ depend only on Δ.

Based on a graph coloring theorem on can show the following corollary, see [19] for details.

Corrollary 6.2. *Any cell $B \in \mathcal{B}_m$ can be guaranteed a fixed fraction $\frac{1}{1+C_{\mathcal{B}}}$ of total time to transmit without interference. This fraction depends only on Δ.*

Scheduling policy

We split the unit of time into slots of length $\frac{1}{1+C_{\mathcal{B}}}$ units and assign one slot for each of the interfering cells $B \in \mathcal{B}_m$. We further split each slot into "down-link" and "up-link" phase of equal duration so that each node within a particular cell $B \in \mathcal{B}_m$ could be transmitting towards the base station in B during the "up-link" phase and be receiving from the base station during the "down-link" phase. In addition, within both the "up-link" and "down-link" phases a base station services the nodes within its Voronoi cell in a round robin fashion.

Remarks on the proposed construction

Our construction for tessellation is similar to that used in [19] and will allow us to argue that the empirical frequencies of some events for sufficiently large numbers m and n with high probability tend to 0. For example, we will be interested in the event for which a particular cell $B \in \mathcal{B}_m$ of the "infrastructure" tessellation contains a number of nodes exceeding some pre specified value. Although, by construction, on average the number of nodes within such a cell

is $n(B) = \Theta\left(\frac{n}{m}\right)$, we still need to ensure that the probability of a "large deviation" from the average is negligible for large enough n, m.

For the set of discs on the plane, the uniform convergence of empirical frequencies of events to the corresponding probabilities has been established in [19]. Then this convergence can be used to argue the same for the cells with the discs-inclusion-containment properties given in Lemma 6.2. The next paragraph summarizes the results which we will be using.

Uniform convergence of empirical frequencies to the respective probabilities for a set of discs on the plane

In what follows we let $\mathcal{D}_r$ denote the collection of all discs D_r of radius r on the plane. Then the following result is proven in [19]:

Proposition 6.4. *Let the points $X_i, i = 1, 2, \ldots, n$ be i.i.d random points on the plane. Define the empirical frequency that a point X_i falls into a disc $D_r \in \mathcal{D}_r$:*

$$F(D_r) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in D_r\}}$$

and let $P(D_r) = \mathbb{P}(X_1 \in D_r)$. Then,

$$\mathbb{P}\left(\sup_{D_r \in \mathcal{D}_r} |F(D_r) - P(D_r)| \leq \varepsilon\right) > 1 - \delta, \tag{6.16}$$

whenever:

$$n > \max\left[\frac{24}{\varepsilon}\log\frac{16e}{\varepsilon}, \frac{4}{\varepsilon}\log\frac{2}{\delta}\right].$$

This proposition follows from the uniform convergence in the weak law of large numbers due to Vapnik and Chervonenkis, and we refer the reader to [19] for the proofs and references therein.

The number of ad hoc nodes contained within an infrastructure cell $B \in \mathcal{B}_m$

Based on Proposition 6.4 one can show the following lemma.

Lemma 6.4. *Let $N_s(B)$ be the number of ad hoc senders falling within an infrastructure cell $B \in \mathcal{B}_m$. Then, in Regime (ii), for some $c_1 > 0$, which is independent of m and n, we have that as $n \to \infty$*

$$\mathbb{P}\left(\max_{B \in \mathcal{B}_m} N_s(B) \leq c_1 \frac{n}{m}\right) \to 1$$

Proof. By Lemma 6.2, each cell $B \in \mathcal{B}_m$ is contained in a disc $D_{r(m)}$ of radius $r(m) = \frac{3}{\sqrt{m}}$. Denote $\tilde{\mathcal{D}}_{r(m)}$ the collection of discs on the plane having such a radius. Then, by Proposition 6.4

we obtain:

$$\mathbb{P}\left(\sup_{D_{r(m)}\in\tilde{\mathcal{D}}_{r(m)}}\left|N_s(D_{r(m)})-nP(D_{r(m)})\right|\leq n\varepsilon(n)\right)>1-\delta(n)\,, \tag{6.17}$$

whenever:

$$n>\max\left[\frac{24}{\varepsilon(n)}\log\frac{16e}{\varepsilon(n)},\frac{4}{\varepsilon(n)}\log\frac{2}{\delta(n)}\right]. \tag{6.18}$$

If we let $\varepsilon(n)=\delta(n)=\frac{50\log n}{n}$, one can show that (6.18) holds. Then, from (6.17):

$$N_s(D_{r(m)})\leq nP(D_{r(m)})+n\varepsilon(n)\,, \tag{6.19}$$

with probability exceeding $1-\delta(n)$ uniformly for all discs $D_{r(m)}\in\tilde{\mathcal{D}}_{r(m)}$. The result follows now from the fact that $P(D_{r(m)})=\frac{3\pi}{m}$ and the fact that $\frac{1}{m}$ is the leading factor in (6.19). □

The number of nodes, choosing intended receivers within an infrastructure cell $B\in\mathcal{B}_m$

Based on Proposition 6.4 one can show the following bound on the number of receivers in a given cell.

Lemma 6.5. *In Regime (ii), the number of ad hoc nodes, $N_r(B)$, that choose their intended receivers in a particular infrastructure cell B is upper bounded for some $c_2>0$, independent of m and n, as:*

$$\mathbb{P}\left(\max_{B\in\mathcal{B}_m}N_r(B)\leq c_2\frac{n}{m}\right)\to 1\,,$$

as $n\to\infty$.

Proof. By definition of $N_r(B)$:

$$N_r(B)=\sum_{i=1}^{n}\mathbf{1}_{\{Y_i\in B\}}\,, \tag{6.20}$$

where Y_i is the node closest to the uniformly and randomly chosen point S_i on D that the node located at X_i chooses to communicate with. Note, that by Theorem 6.1, with probability achieving 1 as $n\to\infty$, each node is at most $2b\sqrt{\frac{\log n}{n}}$ away from a closest to it node, whenever b obeys $2b>\frac{1}{\sqrt{\pi}}$. Therefore, the location S_i must be at most $b\sqrt{\frac{\log n}{n}}$ away from Y_i. This yields, that S_i falls within the set $B\oplus D_{\rho(n)}$, which is the set formed by taking the union of B and all the discs of radius $\rho(n)=b\sqrt{\frac{\log n}{n}}$ centered at the boundary of B. Therefore, by (6.20) it follows that:

$$N_r(B)\leq\sum_{i=1}^{n}\mathbf{1}_{\left\{S_i\in B\oplus D_{\rho(n)}\right\}}\leq\sum_{i=1}^{n}\mathbf{1}_{\left\{S_i\in\bar{D}(B)\oplus D_{\rho(n)}\right\}}\,,$$

where by $\bar{D}(B)$ we denoted the disc of radius $\frac{3}{\sqrt{m}}$ containing the cell B. Applying Proposition 6.4

with $\delta(n) = \varepsilon(n) = \frac{50\log n}{n}$ to the i.i.d. $\{S_i\}_{i=1}^n$ and the disc $\bar{D}(B) \oplus D_{\rho(n)}$, we obtain:

$$\begin{aligned} N_r(B) &\leq n\mathbb{P}\Big(S_1 \in \bar{D}(B) \oplus D_{\rho(n)}\Big) + n\varepsilon(n) \\ &\leq n\pi\left(\tfrac{3}{\sqrt{m}} + \rho(n)\right)^2 + \log n\,, \end{aligned} \tag{6.21}$$

uniformly for all $B \in \mathcal{B}_m$ and with probability approaching 1 as $n \to \infty$. The result now follows by noting that $\frac{3}{\sqrt{m}}$ is the leading factor in (6.21) whenever $m = O\left(\frac{n}{\log n}\right)$. □

The capacity of the scheme

The aggregate bandwidth, available per cell of a base station, by the scheduling policy is $\frac{W}{2(1+C_{\mathcal{B}})}$. Since, by Lemma 6.4, this bandwidth is shared by at most $\frac{2c_1 n}{m}$ nodes for sufficiently large n, each node is guaranteed a rate of

$$\lambda \geq \frac{Wm}{2c_1(1+C_{\mathcal{B}})n}\,, \tag{6.22}$$

via the up-link transmissions.

Now, by Lemma 6.5, the number of nodes, sending to any node within a base station cell $B \in \mathcal{B}_m$ is uniformly bounded as:

$$N_r(B) \leq c_2\frac{n}{m}\,,$$

and hence, the aggregate traffic that a base station has to deliver to the nodes in its cell does not exceed

$$\lambda N_r(B) \leq c_2\lambda\frac{n}{m}\,. \tag{6.23}$$

This shows, that the aggregate down-link traffic at each base station will be accommodated with probability approaching 1 as $n \to \infty$ if:

$$\frac{W}{2(1+C_{\mathcal{B}})} \geq \lambda\frac{n}{m}\,.$$

is feasible. This, combined with (6.22) yields for the achievable throughput:

$$\lambda(m,n) = \Omega\left(\frac{Wm}{n}\right).$$

Thus we have shown the following result.

Proposition 6.5. *The throughput capacity of a random hybrid network under the Protocol Model in Regime (ii) is of order* $\Theta\left(\frac{Wm}{n}\right)$.

6.4.3 Regime (iii).

Note that by Proposition 6.5, whenever $m = \Theta\left(\frac{n}{\log n}\right)$ the throughput of a random hybrid network is of order $\Theta\left(\frac{W}{\log n}\right)$. In the same time, by Proposition 6.3, $\Theta\left(\frac{W}{\log n}\right)$ is the highest possible throughput that could be achieved with any number of arbitrary placed base stations. We thus conclude, that the order of the throughput in this regime is $\Theta\left(\frac{W}{\log n}\right)$ and it could be achieved by deploying only $\Theta\left(\frac{W}{\log n}\right)$ out of total m base stations. Thus we have the following:

Proposition 6.6. *The per user throughput of a random hybrid network under the Protocol Model in Regime (iii) is of order* $\Theta\left(\frac{W}{\log n}\right)$.

6.5 Concluding Remarks.

In this chapter we have investigated the asymptotic per user throughput for a random hybrid network with arbitrarily placed base stations and demonstrated explicit schemes achieving these asymptotic scalings. Our results might be viewed as pessimistic, as they confirm the conclusion of [14] (for two particular routing strategies) that to obtain a significant improvement in capacity for such networks infrastructure investments will need to be high. However, taking another point of view the results are good news since they suggest that when the number of infrastructure nodes exceeds $\sqrt{n/\log n}$, ad hoc nodes will be able to effectively share the spatially distributed infrastructure. In practice the first step towards increasing capacity in wireless networks is to increase the capacity of infrastructure nodes. Thus if their number exceeds $\sqrt{n/\log n}$ we might expect ad hoc nodes that are part of a hybrid network to directly see the benefits of such investments in infrastructure. Additionally, ad hoc relaying of information in a hybrid network can be viewed as an effective way to leverage spatially distributed infrastructure and possibly mobile ad hoc nodes to extend the coverage for power-constrained infrastructure nodes. Thus the cost-benefit analysis of a hybrid network should not be considered simply from the perspective of throughput, but also in terms of the infrastructure cost to service a large, possibly spatially distributed customer base.

Another concern for ad hoc network applications is energy consumption. One might expect the addition of infrastructure nodes to significantly reduce the average energy requirements for transmission and relaying among ad hoc nodes by providing a more efficient communication backbone for traffic that needs to go a long way. In particular, for a large-scale sensor network based on ad hoc wireless nodes and operating under fairly tight energy constraints, the investment in extra infrastructure nodes can pay off handsomely by allowing battery operated sensors to operate over a longer period of time. An unfortunate issue in this context will be the traffic hot spots, and thus increased energy consumption, that ad hoc nodes in close proximity to infrastructure nodes are likely to see. We, however, believe that in reality, better routing

and scheduling algorithms will be able to alleviate these hot spots and increase the throughput scaling by a multiplicative factor.

Finally we have shown a sharp cutoff bound on the throughput that can be achieved in a hybrid network irrespective of the number of infrastructure nodes. Our argument is based on observing that some types of traffic fluctuations can systematically limit the per user throughput. Although this is a direct consequence of the random traffic model we have adopted, it does suggest that pure ad hoc networks may have an unfortunate characteristic. Specifically spatial fluctuations in traffic loads are likely to be difficult to support, unless appropriate infrastructure and routing are provided to quickly enable their dissipation.

6.6 Proof of Proposition 6.1.

We will first introduce a few definitions. Let $\mathcal{P}_m$ denote the set of all possible placement rules for m base stations on the disc D. Let $P_m \in \mathcal{P}_m$ be a particular placement rule, for which $S_k(P_m)$ denotes the position of the k-th base station ($k = 1, \ldots, m$). Let $r(n, P_m)$ denote the smallest common range which ensures that n ad hoc nodes randomly placed on D and m base stations placed according to P_m, will connect into a single cluster with probability 1. Finally, let $r^*(n,m) = \min_{P_m \in \mathcal{P}_m} r(n, P_m)$, and $P^*(n,m) = \arg\min_{P_m \in \mathcal{P}_m} r(n, P_m)$, thus $r^*(n,m)$ is the "connectivity range" under the "best" placement rule $P^*(n,m)$.

We prove the proposition by contradiction. In particular, consider the two scaling regimes:

$$\text{Regime 1}: \quad m = o\left(\tfrac{n}{\log n}\right),$$

$$\text{Regime 2}: \quad m = \Omega\left(\tfrac{n}{\log n}\right).$$

and suppose, in contradiction to the statement of the proposition, that the sequence $\{r^*(n,m)\}_{n=1}^{\infty}$ is such that:

$$\begin{aligned} r^*(n,m) &= o\left(\sqrt{\tfrac{\log n}{n}}\right) \quad \text{in Regime 1}\ , \\ r^*(n,m) &= o\left(\tfrac{1}{\sqrt{m}}\right) \qquad \text{in Regime 2}\,. \end{aligned} \tag{6.24}$$

We let $B_k(n,m)$, for $k = 1, \ldots, m$, denote discs of radius $r^*(n,m)$ centered at $S_k(P^*(n,m))$. Let $U(n,m) = D \cap \left(\cup_{k=1}^{m} B_k(n,m)\right)$ denote the subset of D that is covered by such discs. Clearly, any ad hoc node that belongs to $D \setminus U(n,m)$ can not reach any base station in a single hop. Such nodes would have to connect to other ad hoc nodes in their vicinity. However, we will show below, that under the assumptions in (6.24), the region $D \setminus U(n,m)$ contains ad hoc nodes isolated from any other ad hoc nodes, with positive probability for all sufficiently large n and m.

Let $E_{n,m}(r)$ denote the event that there exists an ad hoc node within $D \setminus U(n,m)$ that

has no ad hoc neighbor within a distance r from it. Let $\rho(n) = \sqrt{\frac{\log n + c}{n}}$, for a fixed $c > 0$. It is easy to check that in both Regime 1 and 2 we have that $r^*(n,m)$ given by (6.24) scales as $o(\rho(n))$. Thus, we have that $\mathbb{P}(E_{n,m}(r^*(n,m))) \geq \mathbb{P}(E_{n,m}(\rho(n))$ for all sufficiently large n and m. Now, note that, by Theorem 2.1 in [90] we have that, for all sufficiently large n, with positive probability that is independent of n, there exists at least one ad hoc node in D which has no ad hoc neighbors within the distance $\rho(n)$ from it. Noting that, by the assumptions (6.24) the area of $U(n,m)$ scales in both Regime 1 and Regime 2 as $o(1)$, the proof of Theorem 2.1 in [90] can be straightforwardly adapted to yield the same conclusion for the nodes within $D \setminus U(n,m)$. Thus, with positive probability independent of n and m, for all sufficiently large n and m there exist ad hoc nodes within $D \setminus U(n,m)$ that have no ad hoc nodes within the distance $\rho(n)$ and, hence, also within the distance $r^*(n,m)$. We thus arrive at a contradiction since we assumed that $r^*(n,m)$ is sufficient to ensure that all nodes are connected.

6.7 Proof of Fact 6.1.

Since the area of each hexagon is $\frac{3\sqrt{3}}{2m}$, there are at most $m_0 \leq \lfloor \frac{2m}{3\sqrt{3}} \rfloor < m$ hexagons fully contained within the unit area disc, from which the first statement follows.

We now introduce a coordinate system on the disc, so that the origin is at the center of D. Consider points on a regular grid P_{ij} with the coordinates $\left(\frac{3i}{\sqrt{m}}, \frac{3j}{\sqrt{m}}\right)$, $i, j = 0, 1, 2, \ldots, \lfloor \frac{1}{6}\sqrt{\frac{m}{\pi}} \rfloor$. Each of those points is within a distance of at least $\frac{3}{\sqrt{m}}$ away from the boundary of the disc. Then for each point P_{ij} there must be a hexagon H_{ij} out of m_0 that is fully contained within D, with its center within the distance of at most $\frac{1}{\sqrt{m}}$ from P_{ij}. Notice that $H_{ij} \cap H_{i'j'} = \emptyset$, since otherwise the points P_{ij} and $P_{i'j'}$ would have to be within the distance of at most $\frac{2}{\sqrt{m}}$. Now, since all H_{ij} are disjoint, the number of them is at least $\left(\lfloor \frac{1}{6}\sqrt{\frac{m}{\pi}} \rfloor\right)^2 = \Theta(m)$. But then $m_0 = \Theta(m)$, which proves the second statement of the lemma.

6.8 Proof of Lemma 6.2.

We first show that each cell $B \in \mathcal{B}_m$ contains a disc of radius $\sqrt{\frac{3}{2m}}$. We can view the hexagonal tessellation $\mathcal{H}_m$ of $\mathbb{R}^2$ as being induced by the seeds placed at the center of each hexagon. By construction, the tessellation $\mathcal{B}_m$ is formed by eliminating some of the seeds of $\mathcal{H}_m$. Since a Voronoi cell of a seed is an intersection of the half spaces associated with points closer to that seed, it follows that this intersection will contain the corresponding hexagon of $\mathcal{H}_m$ when some seeds are eliminated. Now, a hexagon of side $\frac{1}{\sqrt{m}}$ contains a disc of radius $\sqrt{\frac{3}{2m}}$, and hence the first property stated in the lemma holds.

Now let us show that second statement of the lemma, i.e., that each cell $B \in \mathcal{B}_m$ is

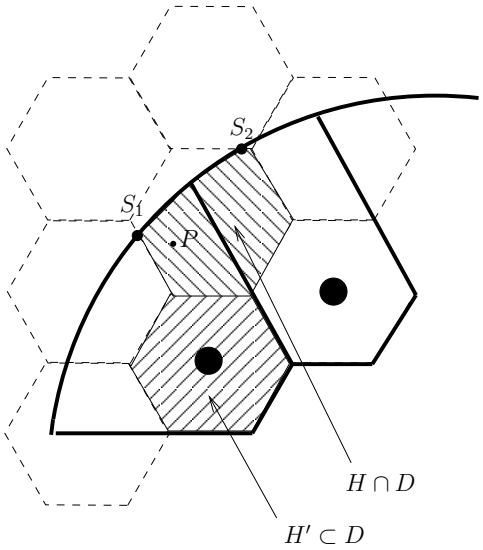


Figure 6.2: H has a neighbor H' fully contained within the disc D.

contained in a disc of radius $\frac{3}{\sqrt{m}}$. This will be true if we can prove that each point P on the disc is within the distance $\frac{3}{\sqrt{m}}$ from one of the base stations. For $P \in H \in \mathcal{H}_m$ and $H \subset D$ (the hexagon is fully contained within the disc) this is obvious since then the distance to the base station centered at the center of the hexagon H is at most $\frac{1}{\sqrt{m}} < \frac{3}{\sqrt{m}}$.

Now consider the case $P \in H \in \mathcal{H}_m$ such that $H \cap D \neq \emptyset$, and $H \not\subset D$. (The hexagon is not fully contained within the disc.) In that case, for sufficiently large m, the boundary of D has to cross the boundary of H in at most two points S_1 and S_2. (S_1 might coincide with S_2, when the circle touches the hexagon at its vertex.)

6.8.1 S_1 and S_2 belong to nonadjacent edges of H

In this case, for a sufficiently large m, there is a hexagon H' fully contained within D and sharing with H an edge (see Figure 6.2). Since the distance from P to the center of H is at most $\frac{1}{\sqrt{m}}$ and the distance between the centers of H and H' is $\frac{2}{\sqrt{m}}$, the point P is within the distance of $\frac{3}{\sqrt{m}}$ from one of the base stations.

6.8.2 S_1 and S_2 belong to adjacent edges of H

There are two cases to consider, one is depicted on Figure 6.3, and the other - on Figure 6.4. In the former case, for a sufficiently large m there is still a hexagon $H' \subset D$ sharing and edge with H, thus the statement of the lemma holds.

In the case of the Figure 6.4, any point P within the intersection $H \cap D$ is also contained within the set G (defined on the picture) for a sufficiently large m. For that m it is easy to see that there is an $H' \in \mathcal{H}_m$ fully contained within D, with its center located within the distance of

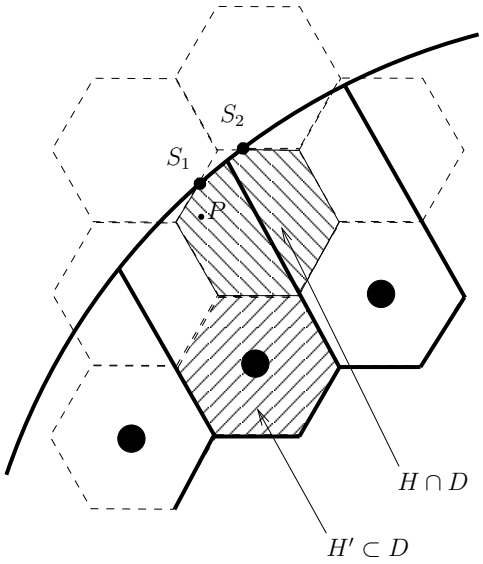


Figure 6.3: H has a neighbor H' fully contained within the disc D.

$\frac{3}{\sqrt{m}}$ from the center of H. Now, since G falls in the ball drawn from the center of H' with radius $\frac{3}{\sqrt{m}}$, we get that the lemma holds for this case too.

6.9 Proof of lemma 6.3.

It is clear that the nodes using a particular base station are located within the range $r_{i.s.}$ from this base station. A node will interfere with another node belonging to a cell $B' \in \mathcal{B}_m$ using the infrastructure communications only if the distance between them is at most $(2+\Delta)r_{i.s.}$. Since the diameters of B and B' are bounded by $r_{i.s.}$, the cell B and all its interfering neighbors B' are located within a common disc of radius:

$$\frac{1}{2}(r_{i.s.} + (2+\Delta)r_{i.s.} + r_{i.s.}) = 2r_{i.s.} + \frac{\Delta}{2}r_{i.s.} = \frac{12+3\Delta}{\sqrt{m}}$$

Since each cell B of $\mathcal{B}_m$ contains a disc of radius $\sqrt{\frac{3}{2m}}$ by construction, B will have at most

$$\left(\frac{\frac{12+3\Delta}{\sqrt{m}}}{\sqrt{\frac{3}{2m}}}\right)^2 \leq C_{\mathcal{B}_m}$$

interfering neighbors.

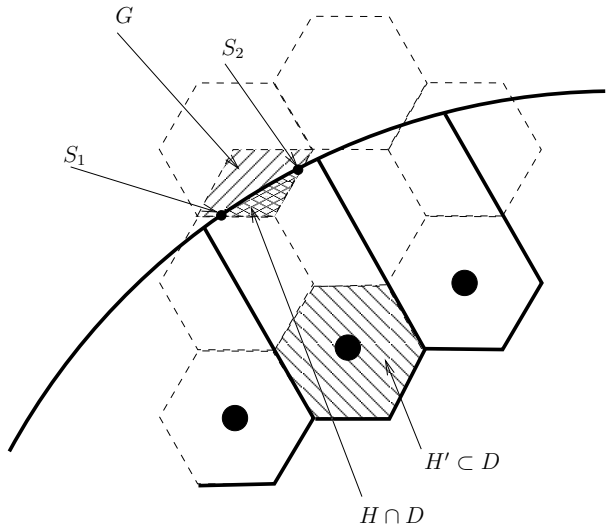


Figure 6.4: Any point $P \in H \cap D$ (shaded region) is within distance of $\frac{3}{\sqrt{m}}$ from the center of H', that is fully contained in D.

Chapter 7

Summary of contributions and possible future directions

In this thesis we have described our latest research towards evaluating wireless systems which may be based on multiple providers using different technologies, and in which end-systems can select among multiple wireless interfaces and/or modes of communication. The specific contributions are as follows. In Part I of the dissertation we have considered a multi-provider network in which users have ability to select between a WAN and a hotspot provider. We have formulated a geometric model for providers spatial interactions and a model for decision-making of individual dual-mode devices. The model is quite general in that it allows for both deterministic and random placement of users and access points and deterministic or random shapes of associated with the access points service zones. We have analyzed the decision-making dynamics and devised a proof of the dynamics' convergence to an equilibrium. We then investigated the decision-making equilibria properties, and classified various equilibria. The "shape of equilibria" in general depends on the realized geometry of service zones, and positions of access points and users. In order to describe the ability of hotspots and the WAN providers to compete with each other we have studied asymptotic properties of equilibria for scenarios in which WAN service zones contain increasingly larger number of users and WLANs. For such scenarios we have shown that the "cutoff" level which describes the largest number of users connected to any of the hotspots in equilibrium converges for sufficiently large WAN service zones to a unique number, which we compared to the typical number of users residing in each hotspot to conclude on competitiveness/viability of the providers.

After that we turned to the analysis of jointly optimal decision-making and backhaul allocation problems which enable performance gains and resource savings in "loosely coupled" multi-provider wireless systems. We have shown that congestion sensitive decision-making is likely to enable significant performance gains, since it would exploit the statistical multiplexing property of the WAN in the most efficient way. We have then devised solutions to backhaul di-

mensioning problems which enable to optimally incorporate the information about users spatial density profiles and WAN rates spatial profiles.

We have also formulated and evaluated decision-making algorithms for "tightly-coupled" multi-provider systems. We started from simple decision-making algorithms akin to the load-balancing algorithms in wired/computer networks literature, but have shown that there are scenarios in which such algorithms perform quite badly since they do not incorporate other-cell interference in decision-making. Then we devised correction terms which allow to incorporate interference in decision-making and bias the decisions appropriately so as to achieve further performance gains.

In Part II we have analyzed the, so-called, hybrid wireless systems, in which users are able to choose between an ad hoc and infrastructure communication modes. We have provided a rigorous analysis of the capacity of hybrid networks and determined the asymptotically (in the number of end-nodes) optimal operation regimes. Our analysis have included the proof for the upper bound on the network capacity that is independent of the routing strategies. We also shown that the upper bound is tight by constructing specific communication schemes in which the capacity order given by the upper bound is achieved.

To conclude, we acknowledge that our thesis presents a series of accomplished studies within the general theme of analysis and design of complex heterogeneous wireless systems. We certainly do not claim that the space of problems which arise within the context of this general area have been exhausted by the ones considered in this dissertation. Thus there is a lot of space for novel research, which could stem off this dissertation. Some of particular promising and challenging directions are outlined below.

- Augmentation of multi-provider systems models:
  - Incorporating different mechanisms for resource allocation and scheduling at wireless access points, see e.g.,[91, 92], and quality of service differentiation.
  - Consideration of more realistic admission schemes, such as schemes with blocking.
  - Design of heterogeneous systems for mixes of traffic, such as, e.g. voice and data, that have different quality of service requirements and thus will have to be assigned different performance metrics.
- Augmentation of hybrid wireless network models:
  - Joint designs of multi-provider and hybrid systems, i.e. systems with both multi-interface diversity and ad-hoc cooperation (e.g. relaying)
- Evaluation of realistic solutions, i.e. WiMax networks combined with WAN networks for better coverage

Bibliography

[1] T.S. Rappaport, *Wireless Communications: Principles and Practice, 2/E*, Prentice Hall PTR, 2002.

[2] T. S. Rappaport, A. Annamalai, R. M. Buehrer, and W. H. Tranter, "Wireless communications: Past events and future perspective," *IEEE Communications Magazine*, pp. 148–161, May 2002.

[3] A. Salmasi and K. Gillhousen, "On the system design aspects of code division multiple access (cdma) applied to digital cellular and personal communications networks," in *IEEE Vehicular Technology Conference*, May 1991, pp. 57–62.

[4] *EIA/TIA/IS-95 mobile station-base station compatibility standard for dual-mode wideband spread spectrum cellular system*, July 1993.

[5] U. Varshney and R. Jain, "Issues in emerging 4g wireless networks," *Computer*, pp. 94–96, June 2001.

[6] GTRAN, "GTRAN dual-mode 802.11/CDMA wireless modem," http://www.gtranwireless.com.

[7] S. Han Y. W. Lee S. Miller L. Salgarelli M. Buddhikot, G. Chandranmenon, "Integration of 802.11 and third-generation wireless data networks," in *Proc. IEEE INFOCOM*, 2003.

[8] CDMA Development Group, "CDMA2000 & Wi-Fi: Making a business case for interoperability," September 2003, http://www.3gnewsroom.com/html/whitepapers/2003/WiFi_Article_9-26-03.pdf.

[9] TROPOS Networks, "Price-performance comparison: 3g vs. tropos metro-scale cellular wi-fi," April 2004, http://www.tropos.com/pdf/price-performance.pdf.

[10] P. S. Henry and H. Luo, "WiFi: What's next?," *IEEE Communications Magazine*, December 2002.

[11] N. Vulic and S. H. de Groot, "Architectural options for WLAN integration into the UMTS radio access level," in *Proc. of the IEEE Vehicular Technology Conference*, May 2004.

[12] G. Rittenhouse, "Next generation wireless networks," in *Proc. INFORMS*, March 2004.

[13] S.V. Hanly and D. Tse, "Power control and capacity of spread-spectrum wireless networks," *Automatica*, vol. 35, no. 12, pp. 1987–2012, December 1999.

[14] B.Liu, Z. Liu, and D. Towsley, "On the capacity of hybrid wireless networks," in *Proc IEEE INFOCOM*, 2003, vol. 3, pp. –.

[15] H.Luo, R. Ramjeey, P. Sinhaz, L. Li, and S. Lu, "UCAN: A unified cellular and adhoc network architecture," in *Proc. MOBICOM' 03*, 2003, pp. 353–367.

[16] W. D. List M. J. Miller and N. H. Vaidya, "A hybrid network implementation to extend infrastructure reach," Tech. Rep., UIUC, Dept. of Electrical and Computer Engineering, January 2003.

[17] J. N. Laneman and G. W. Wornell, "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inform. Theory*, vol. 59, no. 10, pp. 2415–2525, October 2003.

[18] A. Stefanov E. Erkip, A. Sendonaris and B. Aazhang, "Cooperative communication in wireless systems," in *DIMACS Workshop on Network Information Theory*. A.M.S. Publications, DIMACS Volume Series, edited by P. Gupta, G. Kramer and A. J. van Wijngaarden.

[19] P. Gupta and P.R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[20] P. Gupta and P.R. Kumar, "Internets in the sky: The capacity of three dimensional wireless networks," *Communications in Information and Systems*, vol. 1, no. 1, pp. 33–49, 2001.

[21] C. Peraki and S. D. Servetto, "On the scaling laws of wireless networks with directional antennas," in *Proc. 4th ACM MobiHoc*, June 2003.

[22] D. Tse P. Viswanath and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[23] K. Kim J. Chung, C.-S. Hwang and Y. K. Kim, "A random beamforming technique in mimo systems exploiting multi-user diversity," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 848–855, June 2003.

[24] G. J. Foschini, "Layered space-time architecture wireless communication in a fading environment when using multi-element antenna," *Bell Labs. Technical Journal*, pp. 41–59, 1996.

[25] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1451–1458, 1998.

[26] N. Seshadri V. Tarokh, A. Naguib and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criteria in the presence of channel estimation errors, mobility and multiple paths," *IEEE Transactions on Communications*, vol. 47, no. 2, pp. 199–207, February 1999.

[27] 3GPP, "Spatial channel model text description," , no. SCM-134, April 2003.

[28] F. Baccelli, M. Klein, M. Lebourges, and S. Zuyev, "Stochastic geometry and architecture of communication networks," *J. Telecommunication Systems*, vol. 7, pp. 209–227, 1997.

[29] S. Zuyev, P. Desnogues, and H. Rakotoarisoa, "Simulations of large telecommunication networks based on probabilistic modeling," *J. Electronic Imaging*, vol. 6, no. 1, pp. 68–77, 1997.

[30] K. Tchoumatchenko, *Modeling of Communication Networks Using Stochastic Geometry*, Ph.D. thesis, University of Nice - Sophia Antipolis, 1999.

[31] F. Baccelli, B. Blaszczyszyn, and F. Tournois, "Spatial averages of coverage characteristics in large CDMA networks," Tech. Rep. 4196, INRIA (France), 2001.

[32] A. Zemlianov and G. de Veciana, "Modeling competition among wireless service providers," *Tech. Report, UT Austin, available at http://www.ece.utexas.edu/ ˜zemliano/informscomp.pdf*, 2004.

[33] A. Zemlianov and G. de Veciana, "Cooperation and decision-making in a wireless multi-provider setting (extended version)," Tech. Rep., UT-Austin, Dept. of Electrical and Computer Engineering, 2004, http://www.ece.utexas.edu/ ˜zemliano/infocomext.pdf.

[34] M. Benveniste, "Cell selection in two-tier microcellular/macrocellular systems," in *Proc. IEEE GLOBECOM*, 1995, pp. 1532–1536.

[35] K. L. Yeung and S. Nanda, "Channel management in micro/macrocellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 45, pp. 601–612, November 1996.

[36] Mihailescu, X. Lagrange, and D. Zeghlache, "Analysis of a two-layer cellular mobile communication system," in *Proc. of IEEE Vehicular Technology Conf., Phoenix, Arizona*, 1997.

[37] X. Lagrange, "Multitier cell design," *IEEE Communications Magazine*, August 1997.

[38] W. Jolley and R. Warfield, "Modeling and analysis of layered cellular mobile networks," in *Teletraffic Datatraffic in a Period Change*, 1991, vol. ITC-13, pp. 161–166.

[39] R. G. Sheng and L.-S. Tsao, "3G-based access control for 3GPP-WLAN internetworking," in *Proc. of the IEEE Vehicular Technology Conference*, May 2004.

[40] A. K. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS integration for next-generation mobile data networks," *IEEE Wireless Comm.*, vol. 9, pp. 112–124, October 2002.

[41] M. Jaseemuddin, "An architecture for integrating UMTS and 802.11 WLAN networks," July 2003, vol. 2, pp. 716–723.

[42] S. Lincke-Salecker and C. S. Hood, "Integrated networks that overflow speech and data between component networks," *Int. J. Network Mgmt.*, vol. 12, pp. 235–257, 2002.

[43] A. Zemlianov and G. de Veciana, "Cooperation and decision-making in a wireless multi-provider setting," in *Proc. of IEEE INFOCOM*, March 2005.

[44] T. E. Klein and S.-J. Han, "Assignment strategies for mobile data users in hierarchical overlay networks: Performance of optimal and adaptive strategies," *IEEE Journal on Selected Areas in Communication*, vol. 22, no. 5, June 2004.

[45] H. Hotelling, "Stability and competition," *Economic Journal*, vol. 39, pp. 41–57, 1929.

[46] D. Graitson, "Spatial competition a la hotelling: a selective overview," *The Journal of Industrial Economics*, vol. XXXI, no. 1–2, September/December 1982.

[47] E. Ising, "Beitrag zur Theorie des Ferromagnetismus (in German)," *Zeitschr. f. Physik*, vol. 31, pp. 253–258, 1925.

[48] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, revised edition, 1980.

[49] Larry Samuelson, *Evolutionary Games and Equilibrium Selection*, The MIT Press (reprint edition), 1998.

[50] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, MIT Press, 1999.

[51] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.

[52] W. L. Brogan, *Modern Control Theory*, Prentice Hall, 3-rd edition, 1990.

[53] D. Vere-Jones D. Daley, *An Introduction to the Theory of Point Processes*, Springer Verlag, 2 edition, 2002.

[54] F. Baccelli and B. Blaszczyszyn, "On a coverage process ranging from the boolean model to the poisson voronoi tessellation, with applications to wireless communications," *Adv. Appl. Prob.*, vol. 33, no. 2.

[55] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, no. 7, pp. 70–77, 2000.

[56] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM' 03*.

[57] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice Hall, 1989.

[58] E. Altman and L. Wynter, "Equilibrium, games, and pricing in transportation and telecommunication networks," *Networks and Spatial Economics*, vol. 4, 2004.

[59] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Engineers Part II*, pp. 325–378, 1952.

[60] M. Beckmann, C. B. McGuire, and C. B. Winsten, *Studies in the Economics of Transportation*, New Haven: Yale University Press, 1956.

[61] R. W. Rosenthal, "A class of games possessing a pure strategy nash equilibria," *Int. J. Game Theory*, vol. 2, pp. 65–67, 1973.

[62] D. Monderer and L. S. Shapeley, "Potential games," *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996.

[63] CDMA Development Group, "CDMA2000 and WiFi: Making a business case for interoperability," September 2003, http://www.cdg.org.

[64] P. Bremaud F. Baccelli, *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, Springer-Verlag (New York), December 1999.

[65] J.F.C. Kingman, *Poisson Processes*, Clarendon Press, 1997.

[66] K. Ramanan S. Borst, K. Kumaran and P. Whiting, "Queueing models for user-level performance of proportional fair scheduling in wireless data systems," 2002.

[67] H. Kameda, J. Li, C. Kim, and Y. Zhang, *Optimal Load Balancing in Distributed Computer Systems*, New York: Springer-Verlag, 1997.

[68] T. Guven, C. Kommareddy, R. J. La, M. A. Shayman, and B. Bhattacharjee, "Measurement based optimal multi-path routing," in *Proc. of INFOCOM*, 2004.

[69] B. Blaszczyszyn, M. K. Karray, and F. Baccelli, "Blocking rates in large CDMA networks via a spatial Erlang formula," in *Proc. of IEEE INFOCOM*, March 2005, to appear.

[70] V. K. Garg, *Wireless Network Evolution: 2G to 3G*, 2002.

[71] D. P. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1991.

[72] M. Duflo, *Random Iterative Models*, Springer, 2000.

[73] J. W. Cohen and O. J. Boxima, *Boundary Value Problems in Queueing Systems Analysis*, North-Holland Publ. Cy., Amsterdam, 1983.

[74] G. Fayolle and R. Iasnogorodski, "Two coupled processors: the reduction to a Riemann-Hilbert problem," *Z. Wahr. Verw. Geb*, vol. 47, pp. 325–352.

[75] T. Bonald, S. Borst, N. Hegde, and A. Proutière, "Wireless data performance in multicell scenarios," in *Proc. SIGMETRICS/Performance*, 2004.

[76] D. Estrin, R. Govindan, J. Heidmann, and S. Kumar, "Next century challenges: Scalable coordination in sensor networks," in *Proc. MOBICOM '99*, 1999, pp. –.

[77] J.M. Kahn, R.H. Katz, and K.S. Pister, "Next century challenges: Mobile networking for 'smart dust.'," in *Proc. MOBICOM '99*, 1999, pp. –.

[78] I.F. Akyldiz, W. Su, Y. Sankarasubermanian, and E. Cayirici, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, August 2002.

[79] A. Goldsmith and S. Wicker, "Design challenges for energy-constrained ad hoc wireless networks," *IEEE Wireless Communications*, vol. 9, no. 4, 2002.

[80] Timothy J. Shepard, "A channel access scheme for large dense packet radio networks," in *SIGCOMM*, 1996, pp. 219–230.

[81] J. Li, C. Blake, D. De Couto, H.I. Lee, and R. Morris, "Capacity of wireless ad hoc networks," in *Proc. ACM SIGMOBILE '01*, 2001.

[82] D. P. Deed, "How wireless networks scale: The illusion of spectrum scarcity," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, August 2002.

[83] L.L. Xie and P.R. Kumar, "A network information theory for wireless communication: Scaling laws and optimal operation," Submitted to IEEE Transactions on Information Theory, April 12, 2002.

[84] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," *IEEE Transactions on Networking*, vol. 10, no. 4, pp. 477–86, 2002.

[85] N. Bansal and Z. Liu, "Capacity, delay and mobility in wireless ad hoc networks," in *Proc IEEE INFOCOM*, 2003, vol. 3, pp. –.

[86] E. Prevalov and R. Blurn, "Delay limited capacity of ad hoc networks: Asymptotically optimal transmission and relaying strategy," in *Proc IEEE INFOCOM*, 2003, vol. 3, pp. –.

[87] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-off in wireless networks," in *Proc. INFOCOM '04*, 2004.

[88] U.C. Kozat and L. Tassiulas, "Throughput capacity of random ad hoc networks with infrastructure support," in *Proc. of MobiCom*, 2003.

[89] T. H. Cormen and R. L. Rivest, *Introduction to Algorithms*, The MIT Press, 2001.

[90] P. Gupta and P.R. Kumar, *Critical power for asymptotic connectivity in wireless networks*, pp. 547–566, Birkhauser, 1998.

[91] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM' 03*, 2003.

[92] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling for wireless systems with multiple interfaces and multiple constraints," in *Proc. MSWiM*, September 2003.

Vita

Alexander Zemlianov received his M.S. degree in Physics from Moscow State University, Russia, in 1998, and M.S. degree in Electrical and Computer Engineering from the University of Texas at Austin in 2002. His broad research interests are within modeling, control and optimization of stochastic systems.

Before joining the research group of professor Gustavo de Veciana in Summer 2003, Alexander Zemlianov was involved in several interdisciplinary projects. In particular, his Physics M.S. thesis contributed to understanding of anomalous transport in disordered fractal-like structures. In his first years of graduate program at UT he collaborated with UT Business School faculty to design and implement algorithms for hedging financial options. In addition, he worked as a quantitative analyst in two Austin, TX, based companies, "ForwardVue Technologies" and "SciComp, Inc." where he was involved in design and implementation of risk analysis software.

His current research interests include ad hoc and sensor wireless networks, resource allocation in wireless networks, information theory and aspects of physical and MAC layer design in wireless networks.

Permanent Address: 5106 North Lamar, apt. 106
Austin, TX 78751

This dissertation was typeset with LaTeX 2ε[1] by the author.

[1] LaTeX 2ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.