Comparison between Static and Dynamic Modeling Approaches for Heterogeneous Cellular Networks

by

Dariush Fooladivanda

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2014

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In order to accommodate growing traffic demands, next generation cellular networks must become highly heterogeneous to achieve capacity gains. Heterogeneous cellular networks composed of macro base stations and low-power base stations of different types are able to improve spectral efficiency per unit area, and to eliminate coverage holes. In such networks, intelligent user association and resource allocation schemes are needed to achieve gains in performance. We focus on heterogeneous cellular networks that consist of macro and pico BSs, and study the interplay between user association and resource allocation using two modeling approaches, namely a *static modeling approach* and a *dynamic modeling approach*.

Our first study focuses on modeling heterogeneous cellular networks with a *static approach*. We propose a unified static framework to study the interplay of user association and resource allocation under a well-defined set of assumptions. This framework allows us to compare the performance of three resource allocation strategies: *partially Shared deployment*, *orthogonal deployment*, and *co-channel deployment* when the user association is optimized. We have formulated joint optimization problems that are non-linear integer programs which are NP-hard. We have, therefore, developed techniques to obtain upper bounds on the system's performance. We also propose a simple association rule that performs much better than all existing user association rules. We have used these upper bounds as benchmarks to provide many engineering

insights, and to quantify how well different combinations of user association rules and resource allocation schemes perform.

Our second study focuses on modeling heterogeneous cellular networks with a *dynamic modeling approach*. We propose a unified framework to study the interplay of user association, resource allocation, user arrival, and delay. We select three different performance metrics: the highest possible arrival rate, the network average delay, and the delay-constrained maximum throughput, and formulate three different optimal user association problems to optimize our performance metrics. The proposed problems are non-linear integer programs which are hard to solve efficiently. We have developed numerical techniques to compute either the exact solutions or tight lower bounds to these problems. We have used these lower bounds and the exact solutions as benchmarks to provide many engineering insights, and to quantify how well different user association rules and resource allocation schemes perform.

Finally, using our numerical results, we compare the *static and dynamic modeling approaches* to study the robustness of our results. Our numerical results show that engineering insights on the resource allocation schemes drawn out the static study are valid in a dynamic context, and vice versa. However, the engineering insights on user association rules drawn out of the static study are not always consistent with the insights drawn out of the dynamic study.

Acknowledgements

I would like to thank my scientific advisor, Professor Catherine Rosenberg, for giving me such a great opportunity to be part of her research group and to learn a part of her knowledge. I am deeply grateful to my advisor for patient guidance, teaching, advices and dedication to students. Indeed, as a result of her constructive critiques and mentoring, I was able to go along the path of learning to think.

I dedicate this thesis to my parents, Frohar Moalemi and Nasser Fooladivanda, and to my wife, Niloufar Moalemi, for their endless love, support, and encouragement.

Contents

List of Figures

List of Abbreviations

ABSF	Almost Blank Sub-Frame
ASE	Area Spectral Effciency
BS	Base Station
CCD	Co-Channel Deployment
CDMA	Code Division Multiple Access
CIO	Cell Individual Offset
CRE	Cell Range Extension
FDMA	Frequency Division Multiple Access
FFR	Fractional Frequency Reuse
GM	Geometric Mean rate
Hetnet	Heterogeneous network
HSPA	High-Speed Packet Access
LTE	Long Term Evolution
MCS	Modulation and Coding Scheme
MIMO	Multi-Input Multi-Output
NUD	Non-Uniformly Distributed
NUM	Network Utility Maximization

OD	Orthogonal Deployment
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PF	Proportional Fairness
PCF	Picocell First
PS	Processor Sharing
GPS	Generalized Processor Sharing
PSD	Partially Shared Deployement
RA	Resource Allocation
RB	Resource Block
RE	Range Extension
SINR	Signal to Interference plus Noise Ratio
SISO	Single-Input Single-Output
TT	Total Throughput
UA	User Association
UD	Uniformly Distributed

Chapter 1

Introduction

1.1 Overview

Current cellular wireless technologies are mainly based on homogeneous networks, i.e., a set of identical base stations (BS) called macro BSs. In such networks, BSs follow a carefully planned layout and are identical in terms of power levels, antenna configurations, backhaul capacities, etc. BSs are carefully configured to optimize coverage, minimize interference with other BSs, and ensure a roughly equivalent number of users in each cell. A typical homogeneous cellular system is shown in Fig. 1.1 in which a BS is located at the center of each homogeneous cell.

Such networks were able to handle the data traffic generated by customers and managed to be beneficial for both the customers and the operators till the last decade. Over the past decade, telecommunication companies have realized that the volume of data traffic is increasing at a very fast rate, over 100% per year as shown in Fig. 1.2 and Fig. 1.3. Because of this, the operators will face some major challenges in the

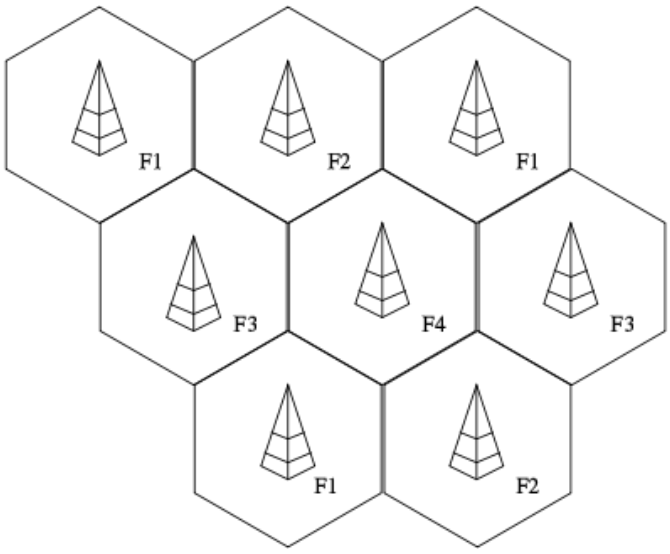
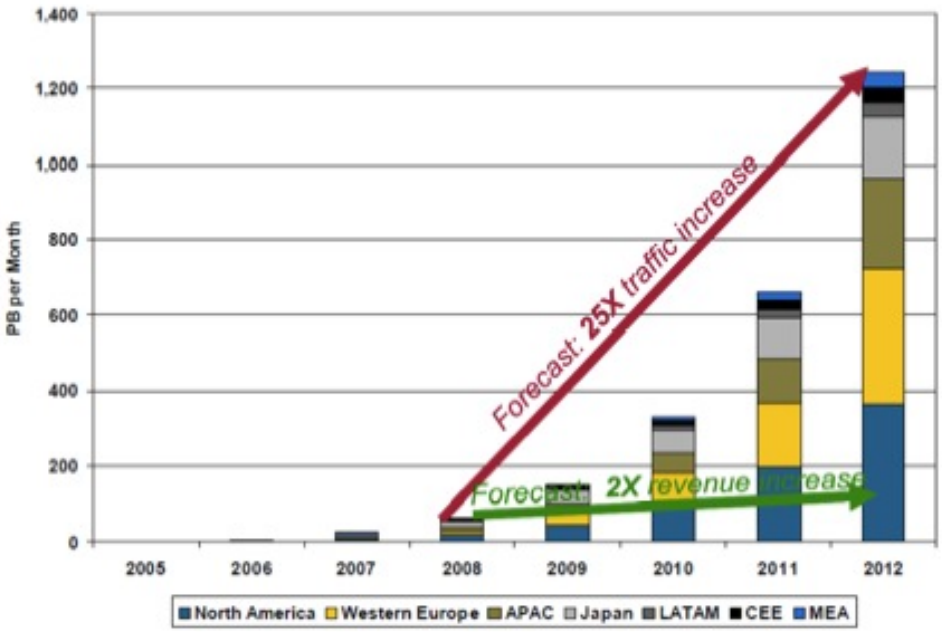


Figure 1.1: Typical Homogeneous Cellular System

near future.

There are some limitations in the amount of data traffic that a homogeneous network can handle. These limitations are mainly determined by the available spectrum and the capacity of the network. The network capacity is determined by information theoretic capacity limits, especially the Shannon-capacity. According to the result in [1], Shannon-capacity for a Single Input Single Output (SISO) cellular system, considering interference as noise, is around 2 bps/Hz spatially averaged over the cell. Hence, the best possible long term throughput is about 2bps/Hz. Although in theory this capacity can be increased by deploying base station cooperation, Multi-Input Multi-Output (MIMO) transmissions, and interference cancelation and alignment schemes, in practice these techniques have not been proven to provide a significant gain due to many practical considerations [2].

Spectrum is the other factor that limits telecommunication companies in handling



Sources: Cisco, from Operators' network data and Analysts, 2008;
Informa, 2008; and Pyramid, "Mobile data revenue will double by 2012,"
Dan Locke, Analyst Insight, 4/2008.

Figure 1.2: Data Traffic Growing Faster Than Revenues

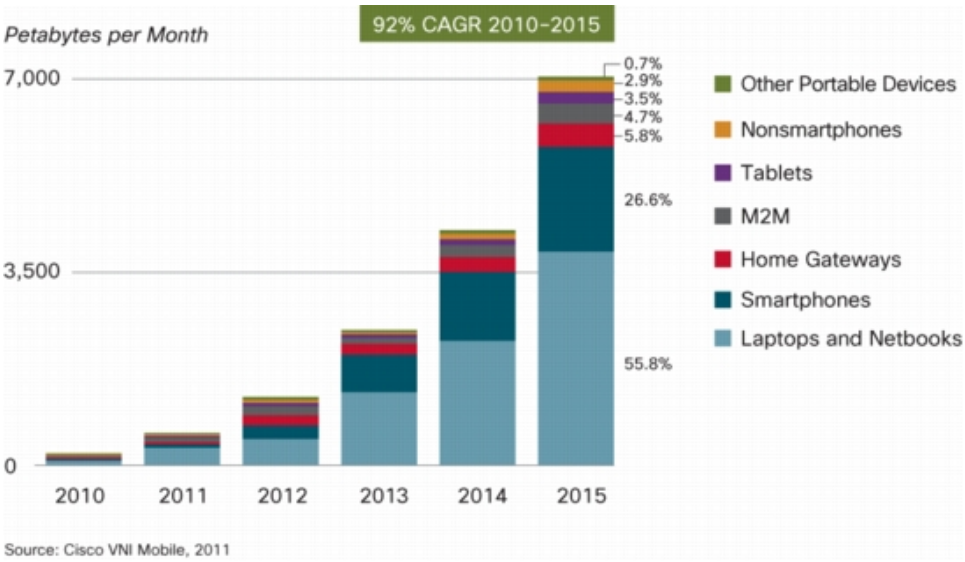


Figure 1.3: Laptops and Smartphones Lead Traffic Growth

high volume of traffic, especially in dense urban areas. Telecommunication companies can either buy more licensed bands or utilize the unlicensed spectrum. Since licensed bands are expensive, network operators prefer to use the available licensed spectrum more efficiently.

Cell size is another factor that affects the number of users that a given BS can handle. By reducing the cell size of the cells (i.e., by adding more BSs) in cellular networks, the network capacity can be improved. Initially, cellular networks were designed with large cell size to keep the infrastructure cost low. Since traffic has increased, operators need to install more BSs to handle higher volume of traffic. This methodology is called cell splitting, and if it is combined with sectorization techniques, it can provide an efficient way of improving network capacity [3], but the cost of infrastructure increases significantly.

While cell splitting can be used to accommodate growing traffic demands, such

an approach can be difficult in dense urban environments because of the challenges in finding new BS sites and the cost of backhauling. In the near future, homogeneous cellular networks will be unable to deliver the required per-user throughput, because a typical modern BS in isolation, employing advanced signal processing, modulation and coding techniques, has already practically reached the information theoretic limits of what an isolated system can achieve. In the future, capacity gains will be achieved by using a wide range of technologies among which will be low-power BSs such as pico, femto, and relay BSs. Cellular networks will become highly heterogeneous, i.e., use different types of technologies to offer higher per-user throughput.

Heterogeneous cellular networks (Hetnets) are composed of macro BSs and low-power BSs of different types, including pico (also called small cells in the literature), femto, and relay BSs. Hetnets are designed to improve spectral efficiency per unit area [4]. The mixture of different BSs with different power levels and different cell sizes can lead to significant gains in performance by offering higher spatial reuse, by eliminating coverage holes, and by creating hot-spots. The LTE-Advanced standard, for example, proposes improvement to network-wide spectral efficiency by employing a mix of low-power BSs [5], [6].

Typically, an operator will place low-power BSs at strategic points to improve performance while keeping the infrastructure cost low. Hence, a user might not always be in the coverage area of a low-power BS. This being said, users should try to associate to low-power BSs if they can, to improve spectral efficiency. This association should improve the user throughput and result in a higher spatial reuse, if resource allocation and interference management mitigate interference among low-power BSs and there are enough resources at the low-power BSs to serve all the users in their

vicinity. Therefore, intelligent user association, resource allocation, and interference management schemes are needed to achieve gains in performance, and the interplay between these schemes has to be studied carefully. There is clearly a complex interplay between the different decisions an operator needs to take during the deployment phase and it is important to perform studies that consider *all* these processes jointly, namely user association, resource allocation and interference management, and scheduling. In the next section, we discuss these schemes in more detail.

1.2 Motivation and Contributions

In this thesis, we study Hetnets on the downlink, and more precisely their *engineering*. At the time of deployment, the operator needs to take many decisions that are a function of the predicted profile of the user population in the region under consideration and the level of service to offer under nominal conditions. Decisions to be taken should include the following processes:

1. **User Association (UA):** This defines a set of rules for assigning users to the different BSs available in the system. A decision to associate a user with one BS will affect the throughput seen by that user, as well as the throughput of the other users associated with all other BSs that would have served that user. In Hetnets, the downlink signal of a macro BSs is typically stronger than that of a low-power BS because of the difference in their transmission powers. Hence, almost all users would associate with macro BSs if user association is based on the downlink received signal strength as it is in homogeneous networks. Therefore, by using the downlink signal strength based association rule, macro

BSs would be overloaded while low-power BSs would serve a small number of users. This would negatively impact the performance of the system, and result in the waste of the resources allocated to the low-power BSs.

In the conventional homogeneous cellular networks, and also in LTE Release-8 [7], [5], user association is based on downlink received signal strength. Many association rules have been proposed that perform better than the conventional rule in Hetnets (e.g., [8], [9]); however, it is not clear which one is the best option since each study is based on a specific resource allocation scheme and a different set of assumptions. In this thesis, we present a benchmark for comparing the performance of the existing association rules in Hetnets in the literature.

2. **Resource Allocation and Interference Management (RA):** Typically, Hetnets are based on OFDM[1], and hence one of the resources to distribute among the different BSs is sub-channels. Another important resource is transmit power. Given a fixed number of channels and a fixed total transmit power (possibly different) at each BS, a RA scheme determines how to allocate the channels among the BSs, and how to use the power budget on the allocated channels at each BS. Hence, in its most complex form, a RA scheme can be seen as a centralized scheduling deciding which BSs should transmit to whom, on which channels, and with what transmit power, at each time. Even in a static scenario where channel gains are known and fixed, and the user association is given, this problem is not tractable due to the very large number of variables.

[1]We assume that the Hetnet as a whole is allocated a frequency band that is divided into M' orthogonal sub-channels (or Resource Blocks (RB) in the context of LTE [10]) where each sub-channel has a bandwidth b. We will use the term channel and sub-channel interchangeably in the thesis.

In its simplest form, a RA scheme might allow each BS to transmit at all time on all sub-channels (and to cope with the resulting interference) using the same power on each channel. In that case, for a given association rule, each BS can locally schedule its own users without the need for any coordination with the other BSs. Clearly, even in this simple case, one expects different performance for different association rules. In this study, *we focus on RA schemes that do not require fine coordination among BSs*, i.e., the schemes determine the number of channels that each BS can use and each BS then uses these channels at all time with the maximum allowed transmit power (distributed over these channels) to schedule its users.

Currently, multiple options exist for managing interference and allocating resources in a Hetnet that includes macro and femto BSs [11], [12], [13], and selecting the right option is a hard problem.

3. **Scheduling Policy**: User throughput is a function of the number of users associated with the same BS as well as the user scheduling policy implemented by the BS and the allocated resource. Hence, the choice of a scheduling policy will significantly impact the system performance. We consider a user scheduling policy based on proportional fairness (PF).

There is a need to develop a *unified framework* to analyze, compare, and evaluate the performance of different resource allocation schemes when user association is either computed optimally or performed via the use of simple rules.

In practice, cellular networks are dynamic systems with respect to users' arrival, service times, user mobility, users' locations, and channel gains. It is hard to study

such systems considering all possible dynamics. However, each of these dynamics can be captured independently. In the literature, several modeling approaches have been proposed to individually capture the dynamic nature of a specific dimension such as users' arrival, service times, users' locations, and channel gains. In the most classic modeling approach (known as *static modeling approach*), we study a snapshot of the system. We assume that BSs' locations are fixed and known, and that there are N *greedy users* placed at random in the system area. These users are greedy in rate, i.e., they want their rate to be as large as possible. We also assume that the BS has an infinite backlog of packets for each of these users. This modeling approach enables us to study the average behavior of many physical layer metrics (such as users' SINRs) over several spatial realizations[2] of the N users in a cellular network.

The *static modeling approach* allows us to formulate many network utility maximization (NUM) problems, and to evaluate the performance of various scheduling, power control, and channel allocation schemes. The main drawback of this approach is that we need to perform multiple network realizations, and to compute the average behavior of our performance metrics over the realizations. *Stochastic geometry modeling* is another static approach in which users are *greedy* in rate, and are placed according to some probability distribution in the system area. This modeling approach allows the study of the average behavior of many physical layer metrics over several spatial realizations of a cellular network without performing multiple network realizations.

[2]A spatial realization corresponds to the random placement of the N users in the system area based on a specific user distribution.

Dynamic modeling approach (also called *flow level modeling approach*) captures the dynamic nature of users' arrival and service times. In this approach, we assume that users arrive in the system, download a file whose size is finite (recall that we focus on the downlink), and depart the system when the file has been received, i.e., we assume a dynamic user population. Such users want to download their files as fast as possible, and want their delay to be as small as possible (i.e., they are sensitive to delay). We fix the power and channel allocations as well as the user scheduling policy per BS (either round robin or processor sharing), and capture the system dynamics by a queueing model which takes into account the users' arrival and departure processes. More precisely, we consider the coverage area of each BS as a queue, and model the cellular network as a set of queues serving the cell area. This approach enables us to study the average behavior of many network layer metrics (such as the average delay in the system) without performing any long-run simulation.

The *dynamic modeling approach* allows us to optimize several delay-based performance metrics when each BS schedules its users locally (using either round robin or processor sharing) given a power control and channel allocation scheme. Unlike the *static modeling approach*, it is hard to evaluate the performance of various user scheduling, power control, and channel allocation schemes using the dynamic approach. Although the *dynamic* approach is less flexible than the *static* approach, it allows us to study the performance of cellular systems in terms of delay-based metrics. *Simulation* is another approach to study the dynamic behavior of an Hetnet. The main drawback of this approach is that we need to perform long-run simulations to evaluate the performance of a user scheduling, power control, and channel allocation scheme. Because of this, it is hard to evaluate the performance of many combina-

Table 1.1: Comparison between the static and dynamic modeling approaches.

Characteristics	**Static Approach**	**Dynamic Approach**
User population	Static with fixed number of users	Dynamic
Number of users	Fixed irrespective of the deployed UA and RA	Highly depends on the deployed UA and RA
File size	Infinite	Finite
Performance metric	Throughput	Delay and Throughput

tions of user scheduling, power control, and channel allocation schemes. Although this approach allows us to study the dynamic behavior of an Hetnet, it makes the optimization of our performance metrics very difficult.

As mentioned above, each of the developed modeling approaches captures different aspects of cellular systems, and models the system under a specific set of assumptions (see Table 1.1). When we model cellular networks under different sets of assumptions (e.g., static or dynamic user population), we might get different engineering insights. Engineering insights that are not "robust" to the set of assumptions are somewhat dubious. Real cellular networks are dynamic systems with users joining and leaving the system. It is not clear whether a combination of a resource allocation and user association scheme that is performing well for a static user population, can also perform well in a dynamic system.

In this thesis, we focus on Hetnets that consist of macro and pico BSs, and study the interplay of user association and resource allocation. We first study Hetnets with a static approach and then with a dynamic approach. We study the interplay of the network processes (i.e., user association and resource allocation), and provide many engineering insights for a static and dynamic user population. The purpose of this study is twofold: First, it is to study the interplay between the various options that a cellular operator has to choose from, and second, to compare the static and dynamic

modeling approaches to make sure that the conclusions that can be drawn out of the static study are valid in a more dynamic context.

Modeling Hetnets using a static approach: We study a snapshot of the system on the downlink, and assume that there are N *greedy users* in the system, i.e., a static user population. We define precisely the Hetnet that we study in terms of the processes (i.e., user association and resource allocation), and formulate a "one-shot" joint user association and resource allocation problem. Our framework is centralized and static since we consider a snapshot of the system both in terms of user deployment and channel gains. This framework allows us to perform an offline study of different combinations of resource allocation and user association schemes, and to select the best performing ones. Our main contributions on the static modeling approach can be summarized as follows:

1. We formulate a centralized static unified framework to analyze and compare several combinations of association rules and RA schemes. We consider three RA schemes: *co-channel deployment* (*CCD*), *orthogonal deployment* (*OD*), and *partially shared deployment* (*PSD*). For *CCD*, we formulate an optimal user association problem. For *OD* and *PSD*, we formulate an optimal joint user association and resource allocation problem. For these three problems, we consider an objective function corresponding to PF among all users in the system (we call this global PF). These three problems are multi-purpose in that, they can be used to compute the optimal association for each RA scheme under consideration (along with the optimal channel allocation for *OD* and *PSD*), which gives us a benchmark (i.e., an upper bound) on the performance to be expected for each scheme. In the case of *CCD*, the problem can be used to

compute the performance of a given association rule. In the case of *OD* and *PSD*, the problem can be used to compute, for a given association rule, the optimal splitting and the corresponding performance.

2. We show how the global PF criteria yields a solution in which each BS schedules its users using local PF.

3. Although the problems in their more general form are non-linear integer programs, we are able to develop numerical techniques to compute tight upper bounds on the performance for small to large systems.

4. We use the numerical results to compare the three RA schemes when the association is optimal. We find that under our assumptions (especially the one on the absence of coordination among BSs), *OD* and *PSD* perform significantly better than *CCD*.

5. We then focus on *PSD* and *OD*, and study the impact of different parameters and how different simple association rules perform. In particular, we propose a simple association rule and show that it works better than the existing association rules.

Modeling Hetnets using a dynamic approach: We consider dynamics in users' arrival and service times, i.e., we assume that users arrive in the system, stay while downloading the file they need, and then depart. We capture the system dynamics by a queueing model which takes into account the users' arrival and departure processes. More precisely, we consider the coverage area of each BS as a multi-class

processor-sharing queue, and model the Hetnet as a set of generalized processor sharing (GPS) queues serving the cell area.

There is a complex interplay between user association, resource allocation, user arrival, and delay in Hetnets. To precisely understand such an important interplay, we focus on the long-run performance of the Hetnet, i.e., the set of GPS queues, and define the Hetnet that we study in terms of user association, resource allocation, and user arrival. We select three different performance metrics: the highest possible arrival rate, the network delay, and the delay-constrained maximum throughput. We then formulate three different optimal user association problems to optimize these metrics. The proposed framework allows us to study Hetnets in a more dynamic context, i.e., to choose the right combination of resource allocation and user association schemes so that a set of delay-based metrics is optimized. Our contributions on the dynamic modeling approach can be summarized as follows:

1. We formulate a centralized framework to analyze and compare different combinations of UA and RA schemes in Hetnets with a dynamic user population. We consider the three RA schemes *CCD*, *OD*, and *PSD*, and three different objective functions corresponding to each of the delay-based performance metrics, namely the highest possible arrival rate, the network delay, and the delay-constrained maximum throughput. Given a resource allocation and its parameters, we formulate three optimal user association problems to optimize the objective functions. The proposed problems enable us to analyze and compare different combinations of UA and RA schemes in Hetnets with a dynamic user population.

2. The proposed user association problems are non-linear integer programs. We develop numerical techniques to compute either the exact solutions or tight lower bounds to these problems.

3. We compare the three RA schemes in terms of the highest possible user arrival rate. We find that *OD* and *PSD* perform significantly better than *CCD*.

4. We then focus on *PSD*, and study the delay performance of different association rules as well as the impact of different system parameters. In particular, we show that the proposed association rule is performing better than the existing association rules.

Finally, we use the proposed frameworks (the static and dynamic ones) to compare the two modeling approaches to make sure that the conclusions that were drawn out of the static study are valid in a more dynamic context. This study is a first step to systematically compare different user association rules and resource allocation schemes for Hetnets, and to compare the two modeling approaches extensively used to model Hetnets. It shows the critical impact of the association rule and the resource allocation scheme as well as the modeling approach in achieving good performance.

1.3 Outline

The rest of the thesis is organized as follows: Chapter 2 presents an overview of the related works. In Chapter 3, we focus on a Hetnet with a static user population. We describe the system model, and formulate a joint user association and resource allocation problem for a static user population. We also provide numerical results

along with engineering insights. In Chapter 4, we consider a Hetnet in which users arrive into the system, download a file, and depart the system. We focus on the long-run performance of the Hetnet, and define the Hetnet that we study in terms of UA, RA, and user arrival. We then formulate three different optimal user association problems to optimize the highest possible arrival rate, the network delay, and the delay-constrained maximum throughput. We also provide numerical results along with engineering insights. In Chapter 5, we compare the two modeling approaches (i.e., the static and the dynamic ones) to make sure that the conclusions that were drawn out of the static study in Chapter 3 are valid in a more dynamic context. Chapter 6 concludes this thesis by summarizing all the insights drawn out of this thesis.

Chapter 2

Literature Background

Resource allocation, interference management, and user association (known as cell-site selection) are well known problems in the area of wireless networking. They have a direct impact on each other. Because of this, some researchers have worked on joint resource allocation, interference management, and user association schemes. In this chapter, an overview of the state of the art for these problems is provided that includes the previous works in the context of both homogeneous and heterogeneous cellular networks. In the following, we mainly focus on frequency division multiple access (FDMA) systems. However, some examples of code division multiple access (CDMA) Hetnets are also provided.

2.1 Resource Allocation and Interference Management

There are different interference management techniques such as power control, channel allocation, interference cancellation (successive or parallel) [14], and interference alignment [15] in the literature. In this section, we focus on interference management via channel allocation among BSs.

2.1.1 Homogeneous Networks

Extensive work has been done on resource allocation and interference management in FDMA homogeneous networks (i.e., that include only macrocells). In FDMA systems, the intra-cell interference is negligible, since we can assume intra-cell users are orthogonal to each other. Hence, the main source of interference is inter-cell interference. Since co-channel interference between adjacent cells can be very large, in traditional FDMA cellular networks to avoid interference among co-channel cells, co-channels cells are spaced a couple of cells away [16]. This technique is called "Channel reuse", and the required minimum spacing called "Reuse Distance" is determined by the minimum required average signal to interference plus noise ratio (SINR) in the system. "Reuse Distance" is defined as the distance between the centers of the cells that use the same channels.

Channel reuse plays an important role in the performance of conventional cellular systems, since it determines the amount of interference between co-channel macrocells [16]. Planning of conventional cellular systems has been based on channel reuse methods. However, in recent years, more sophisticated channel allocation schemes

have been proposed that can utilize frequency resources more efficiently.

In cellular systems, since all BSs transmit with equal power, it is reasonable for the user to associate with the BS that provides the strongest signal (the best SINR). In such scenarios, considering the high traffic aggregation, "reuse factor one" in which all BSs have access to the available frequency band can be an appropriate resource allocation scheme for cell-interior users within a cell [17], but for cell-edge users "reuse factor one" can lead to high interference. In a given cell, the users who are closer to the BS of the cell rather than the neighboring cells are called cell-interior users, and users who are at the edge of the cell are called cell-edge users.

Fractional frequency reuse (FFR) [18] has been proposed as a solution to mitigate the inter-cell interference for both cell-edge and cell-interior users while using frequency resources more efficiently. In FFR systems, the bandwidth is divided into two parts corresponding to cell-interior and cell-edge users. For cell-interior users, "reuse factor one" can be the best bandwidth allocation, since these users are far enough from the neighboring BSs. In contrarily, cell-edge users are close to the neighboring BSs; and so "reuse factor one" is not the right reuse for cell-edge users, since it results in high interference for cell-edge users, and it degrades the rate of cell-edge users. Because of this, typically a reuse factor that is greater than one (e.g., reuse factor 3) has been used for cell-edge users to avoid inter-cell interference between the neighboring BSs.

In FFR cellular systems, there are some tradeoffs between different system performance metrics such as rate improvement, coverage for cell-edge users, overall network throughput, and spectral efficiency, and hence, the design of FFR systems has a direct impact on various performance metrics of the system in hand [19]. For example,

in [20], [21] and [22], the authors focus on the optimal design of FFR systems to maximize the network throughput while in [23] the spectral efficiency is considered as the objective.

2.1.2 Heterogeneous Networks

In Hetnets that include macro and pico BSs, the throughput of the system can be degraded by co-channel interference among neighboring cells. In [24], Yonezawa *et al.* propose a central channel allocation algorithm to minimize interference among co-channel picocells assuming user association and channel allocation among pico and macro BSs are given. In this algorithm, a central node collects some information from users to construct an interference graph. Then, by using the chromatic polynomial in graph theory, the central node determines whether the graph can be colored by the given number of channels or not. If the answer is "No", then the algorithm prunes an edge that means ignoring the interference constraint and increasing the level of interference in the network, and then tries to color the graph again. The final channel allocation minimizes the interference level between co-channel picocells for a given number of channels. Some simulations and experiments are done to demonstrate the effectiveness of the algorithm.

The same problem is addressed in [25] and [26]. The authors propose an algorithm in which users send the interference information to a central node. Based on the received information, the central node assigns sub-channels to pico BSs recursively until interference is higher than a given interference level. Although the authors show the effectiveness of their proposed algorithms based on some computer simulations and experiments, there are numerous drawbacks with this algorithm that diminish

the performance of the system. For example, the performance of the algorithm could be diminished if the number of available sub-channels is relatively small as compared to the cell density. Moreover, the algorithm could make the network to oscillate in terms of channel assignment, since it might change the frequency sub-channels of pico BSs after receiving the feedback information.

In [27], Chandrasekhar *et al.* study the problem of resource allocation in FDMA based Hetnets that include macro and femto BSs. The authors propose an optimal decentralized algorithm for channel allocation between femtcells and the macrocell BS (in a given cell) that maximizes the area spectral efficiency (ASE) while both macrocell and femtocell users can attain a minimum predetermined data rate. The algorithm is analyzed and its performance is demonstrated based on some simulation results. The numerical results show that channel allocation between the macrocell and femtocells depends on the minimum data rate requirement of users, the hotspot density, and the co-channel interference between macrocells and femtocells. In this work, it is assumed that femtocells pick their frequency sub-channels randomly. This assumption is very strong, since typically random channel assignment does not lead to significant gains in throughput [28]. All derived expressions are based on this assumption. Therefore, the provided optimum frequency division expressions and evaluations cannot be considered as a general result.

In [28], the problem of sub-channel assignment among femtocells in Orthogonal Frequency Division Multiple Access (OFDMA) femtocell networks is addressed. Two sub-channel assignment algorithms are proposed that enable femtocells to reduce inter-cell interference and improve the network capacity. The algorithms operate in two phases called “sensing phase” and “tuning phase”. In the sensing phase,

femtocells and users measure interference in all sub-channels and then broadcast the measurements to other femtocells. In the tuning phase, based on the received reports, each femtocell adjusts its sub-channels such that interference is minimized in the network. Some simulation results are provided that show the effectiveness of the algorithm in reducing interference in the network. Although the authors impose some conditions on the "tuning phase" to prevent possible oscillations in terms of channel assignment, there is no guarantee that the algorithm does not oscillate and it converges to an optimum (or sub-optimum) channel assignment.

In Hetnets where the spectrum is shared (e.g., CDMA femtocells) and there is no coordination among BSs in different tiers, cross-tier interference limits the network capacity. In [29], Chandrasekhar *et al.* derive a fundamental relation that provide the largest macrocell SINR given any set of feasible femtocell SINRs, and they propose a distributed utility-based SINR adaptation at femtocells that mitigates cross-tier interference at macrocells caused by femtocells. In this algorithm, each femtocell maximizes its individual utility that consists of a SINR based reward and a cost corresponding to the generated interference by the femtocell. The provided numerical results show more than 30% improvement in mean femtocell SINRs relative to Foschini-Miljanic algorithm [30].

Three resource allocation options, including *co-channel deployment* (*CCD*), *orthogonal deployment* (*OD*), and *partially shared deployment* (*PSD*), have been proposed in 3GPP to share resources between macro and pico tiers [31]. In *OD*, low-power BSs that have the same transmit power share the same set of resources. This solution mitigates interference among different classes of BSs since BSs transmitting on the same sub-channels have the same transmit power. In *PSD*, a set of frequency resources

is dedicated to macro BSs, and the rest of the available frequency resources is shared among the macro and low-power BSs. In this solution, capacity gains can be achieved by using low-power BSs without affecting the coverage of the macro BSs. Note that to reduce interference among macro and low-power BSs on the shared spectrum, macro BSs transmit with reduced power on the shared frequency resources. This solution provides an efficient way of using resources without affecting coverage area of macro BSs.

In *CCD*, all BSs have access to the whole set of channels. This solution is considered more efficient for systems with limited spectrum since it avoids spectrum partitioning, and for systems in which *PSD* is not supported by user equipments. *CCD* results in high interference among macro and low-power BSs so that the coverage of low-power BSs is reduced and their capacity gains are diminished if no interference management technique is deployed in the system. To mitigate the interference among co-channel BSs, 3GPP introduced *almost blank subframe* (ABS), a subframe in which the macro tier is not allowed to transmit data [31], [32]. Low-power BSs can schedule their users who are close to a macro BS, in the allocated ABSs.

In [33], Cierny *et al.* use stochastic geometry to find the minimum number of ABSs so that users' throughput requirements are satisfied. The authors propose a semi-analytical formula which can be used to evaluate the performance of Hetnets comprising macro/femto BSs and macro/pico BSs. Their numerical results show that in Hetnets comprising macro/femto BSs, the white residue interference in ABSs is tolerable for femto users while in Hetnets comprising pico/macro BSs the white residue interference in ABSs significantly affects the required number of ABSs

In [34], the authors explore the performance of *OD* and *CCD* with the assumption

that the system uses the conventional association rule in which a user associates with the BS that provides the highest downlink signal power. The authors show, via simulation, that *CCD* achieves a higher average throughput when the channels are equally divided between the macro and pico BSs.

In summary, in the context of Hetnets (multi-tier networks), there are two main questions related to interference management via channel allocation:

1. How many sub-channels should be allocated to a tier in multi-tier networks ?
2. Assuming channel allocation between tiers is done, what is the best channel assignment to bases stations in each tier ?

Researchers have tried to find reasonable answers to these questions for pico and femtocell networks with a *static user population*, but the proposed solutions are not general enough. In general, it is hard to answers these questions using different fading models (especially in the *dynamic modeling approach*). In this thesis, we will assume that fast fading is averaged out at the link level, and the sub-channel gains account for the path loss and shadowing effects. We then try to answer these questions for Hetnets that include pico and macro BSs, using the *static* and *dynamic* modeling approaches.

2.2 User Association

2.2.1 User Association in Homogeneous Networks

There is extensive work on user association focusing on different objectives such as load balancing, decreasing call blocking probability, and increasing the number of

connected users. In [35], the problem of joint power allocation and user association in spread spectrum cellular networks is formulated, and a distributed power control and user association algorithm that achieves capacity of wireless spread spectrum systems is provided. Based on some numerical examples, it is shown that the algorithm effectively associates users with different BSs such that interference is minimized.

In High Speed Packet Access (HSPA) cellular networks, load balancing in terms of distributing users among different BSs can result in smaller blocking rate for voice traffic, higher throughput and better minimum rate. In [36] and [37], the problem of load balancing in HSPA networks is addressed. The authors propose a load-aware handoff/cell-site selection scheme that maximizes the number of connected users, reduces the number of hot-spots by doing load balancing among different BSs, and increases the robustness of the system to asymmetric load dynamics across the network.

The work in [38] deals with the planning problem of frequency allocation in FDMA voice systems given the locations of BSs. The authors try to maximize the number of connected users in the system given a certain blocking probability. They model the interplay between user association and frequency allocation as a linear integer program subject to some constraints on the number of channels determined by Erlang B-formula as the minimum number of channels required to keep blocking probability below the given threshold. The authors do not provide any numerical result for the proposed benchmark. Only a short discussion on the complexity of the proposed problem is provided in the paper.

In [39], Klein *et al.* studied the cell selection problem for mobile users considering the effect of user's velocity and the amount of data to be transmitted. Based on

the user's velocity and the amount of data to be transmitted, two user association schemes are proposed, and some analytical expressions for the optimal thresholds on the velocity and the amount of data to be transmitted that minimize the expected system load, are provided. The results are general, and hold for any user distribution and call arrival rate. In addition, some on-line user association schemes that do not require a priori information of the network parameters and achieve near optimal performance are also provided.

2.2.2 User Association in Heterogeneous Networks

In [40], the cell selection problem for two tier cellular networks that include macro and micro BSs is studied, and a user cell selection algorithm that reduces the total uplink transmit power required to achieve a given target SINR is proposed. In this algorithm, users will be divided into different set of groups, and then users in each group compete with each other to choose the serving BS that requires the smallest transmit power to achieve the target SINR. Based on some simulation results, it is shown that the algorithm can improve the performance in terms of the required power at the BSs, the transmit power of users, and the load balancing among cells compared with the conventional user association scheme.

In [41], Yang *et al.* consider hand-off between different technologies in heterogeneous wideband code division multiple access (W-CDMA) networks as a multi-dimensional problem. The authors propose a multi-dimensional adaptive SINR based vertical hand-off algorithm (MASVH) that combines the effects of SINR, user required bandwidth, user traffic cost, and BS utilization to make handoff decisions. Some performance evaluations are done that demonstrate the effectiveness of the algorithm in

improving throughput, and reducing dropping probability.

In [8], the authors consider the effect of user association on the network's throughput for a fixed partitioning of resources between the macrocell and some picocells. A simple association rule called "Range Extension" (RE) is proposed, and the authors show by simulation that it can improve the network's throughput as compared to an association based on SINR. In RE, users associate with the BS with the minimum path loss rather than the BS with the maximum downlink SINR. The performance of RE is demonstrated via simulation in which there are less than 50 users. Clearly, this limited setup is not suitable for demonstrating the effectiveness of the association rule. There are some system parameters such as the number of channels allocated to picocells, user population, and their distribution that can affect the performance of RE, and there are some scenarios in which SINR based association performs much better than RE.

In [42], a new association rule called "Range Expansion" is proposed. It adds a bias to the reference signal's received power from pico BSs to artificially extend their coverage [31]. In 3GPP, this association rule and its parameter are called "Cell Range Extension" (CRE) and "Cell Individual Offset" (CIO), respectively [32]. The bias can be selected such that users associate with the BS with the minimum path loss, i.e., RE [8]. In [43], Mukherjee *et al.* study the impact of range expansion bias on the distribution of users' SINRs and the spectral efficiency in Hetnets comprising macro and pico BSs. Using stochastic geometry models, the authors derive closed-form expressions for the spectral efficiency of the Hetnet as well as the SINR distribution of users associated with macro and pico BSs.

In femtocell networks, two types of access methods called "Open Access" and

"Closed Access" have been defined for femtocells [44]. In "Closed Access", only administered users can connect to a femtocell, but in "Open Access" all users who can hear a given femtocell are allowed to be connected to its BS. Clearly, "Closed Access" limits the number of users connected to femtocells while "Open Access" does not impose any restriction on users to connect to any femtocell.

In [45], Chandrasekhar *et al.* analyze "Open Access" and "Closed Access" for the downlink transmission in Hetnets that include macro and femto BSs. It is assumed that the system uses round robin scheduling. The authors provide some expressions for the SINR distribution of various areas within a cell, and based on these expressions the average sum throughput of femtocell and macrocell users under both open and closed access methods have been determined. It is shown that femtocell users prefer "Closed Access" in terms of having higher throughput while macrocell users prefer "Open Access" especially at the cell edge.

In [46], Son *et al.* study the the problem of optimal user association in multi-cell networks. The authors formulate the proportional fair association problem, and propose an offline and online user association schemes. In the proposed online algorithm, users send their information to their current BSs, and BSs send their users' information to a central node. The central node finds which user(s) is the best candidate for re-association. The central controller allows only few users to associate or re-associate at each step. The proposed online algorithm uses a notion of average throughput as the decision making metric, instead of the signal strength in conventional systems. The authors numerically show that the online algorithm works significantly better than the conventional association rule.

In [47], Kim *et al.* consider a dynamic system where users arrive randomly into

the system, download files, and leave after being served. The authors propose a general framework for user association in wireless networks, and capture the dynamics of the system with a flow-level queuing model. The proposed framework is general and applicable to Hetnets since there is no limitation on the transmit power of BSs. The authors formulate the user association problem as a convex optimization problem. They propose and analyze an iterative distributed user association policy that converges to a global optimum and tracks slowly varying traffic loads. The proposed algorithm supports a family of load-balancing objectives such as rate-optimal, throughput-optimal, and delay-optimal objectives. Finally, they propose admission control policies for the scenario where the system is overloaded and cannot be stabilized. It is shown that the optimal admission control policy blocks all flows at cells edges.

In summary, the performance of an association rule depends on different parameters such as transmission power, number of available sub-channels, user density, user population behavior (i.e., static or dynamic user population), scheduling policy, and access methods. In Hetnets, the interplay between these parameters becomes more complicated, since different tiers have different transmission powers, cell sizes, and different access methods. As mentioned above, recently, some researchers have worked on user association in Hetnets, but with many simplifying assumptions. In this study, we find the optimal association, and propose a simple association rule. We also compare the performance of different association rules under different system parameters in terms of number of users, user population (static or dynamic) and density, and number of allocated sub-channels.

2.3 Joint Resource Allocation and User Association

In [48], Madan *et al.* study the effect of joint user association and resource allocation among macro and pico BSs for a static user population. An OFDM system is considered in which the total bandwidth is divided into M sub-channels. Power and scheduling time (on a per sub-channel basis) are the resources that are allocated among different BSs. A global high level scheduling problem is formulated to maximize at each time-slot the sum of the logarithm of the rates as a function of several variables, including power levels, scheduling, sub-channels, and user association. However, since this problem is a very large combinatorial problem, they propose heuristic algorithms for adaptive user association and resource partitioning. The performance of these heuristic algorithms is demonstrated via simulations. It is shown that the heuristics can achieve near optimal performance at least for small networks.

In [48], Madan *et al.* compare the performance of different user association and resource allocation schemes. To do this, they consider a system composed of 2 macro and 10 pico BSs, and 20 users, and compare the performance of the conventional user association rule and "Rang Extension" under different resource allocation schemes and different user utility functions. The resource allocation schemes are "reuse one" (both macro and pico BSs have access to the same set of sub-channels) and "fixed resource partitioning" (the resources are allocated equally between the macrocells and the picocells). The numerical results show that "Range Extension" associates more users with low-power BSs, and significant performance gains can be achieved by using such simple user association rules.

In [49], the performances of "Range Extension" and the conventional association rule are compared under a specific frequency allocation in which the resources are equally divided between the macroells and the picocells. The authors show via simulation that the number of users served by picocells is already large enough with the conventional user association rule, and "Range Extension" does not improve users' throughput significantly. In [9], Tongwei *et al.* propose a new user association rule called "Based on Queue (BQ)". In BQ, a user associates with the low-power BS that has the strongest signal among other low-power BSs and its signal power is greater than the strongest downlink macrocell signal minus 25 dB. BQ associates more users with picocells. The authors compare BQ's performance under two different resource allocation schemes called overlap inter-cell interference coordination (Overlap ICIC) and non-verlap inter-cell interference coordination (Non-overlap ICIC). In "Overlap ICIC", macrocell BSs use half of the available frequency while picocells can use the entire frequency band. In "Non-overlap ICIC", the available bandwidth is equally divided between macrocells and the picocells. Finally, it is shown via simulation that the new scheme works better than "Range Extension" and the conventional user association rule.

In [50], Fooladivanda *et al.* develop a *unified framework* to analyze, compare, and evaluate the performance of different user association rules in Hetnets using max-min scheduling for *Orthogonal deployment*. The authors formulate a joint user association and resource allocation problem, and propose a simple association rule called "Picocell First". The proposed problem is an integer linear program which can be solved for relatively large networks. The authors obtain exact solutions to the proposed problem, and numerically show that the proposed association rule can

improve the minimum throughput as compared to an association based on SINR. In the next chapter, more details are provided on the performance of "Picocell First".

As mentioned in this section, there are some challenging interplay in Hetnets, e.g., between the user association and the channel allocation, that have a significant impact on the system performance. Extensive work has been done on user association and resource allocation schemes in Hetnets for a static user population, but none of these works can be used as a unified benchmark to compare the performance of existing user association rules and resource allocation schemes. It is not clear why some simple association rules perform well in one scenario while these rules do not work well in another scenario. Moreover, none of the existing works explore the interplay between user association and resource allocation in a dynamic context. As mentioned earlier, real cellular networks are dynamic systems with users joining and leaving the system. It is not clear whether a combination of a resource allocation and user association scheme that is performing well for a static user population, can also perform well in a dynamic system. In this thesis, we try to find some reasonable answers to these challenging questions.

Chapter 3

Hetnets with a Static User Population

3.1 System Model

We consider a multi-tier communication system composed of several macro BSs. Each macro BS is overlaid with B pico BSs that are identical in terms of transmit power, antenna gain, and backhaul capacity. We focus on the macro BS at the center (cell 0). The coverage area[1] of the macro BS is covered by a grid corresponding to possible locations for the users (i.e., we discretize the set of locations at which users can be). Let $\mathcal{B}$ and $\mathcal{L}$ denote the set of pico BSs deployed by the operator ($|\mathcal{B}| = B$) in the coverage area of the macro BS and the set of possible user locations within the cell area ($|\mathcal{L}| = L$), respectively. We study the downlink and make the following

[1]The coverage area of a macro BS is the area in the multi-tier communication system without pico BSs (a homogenous system) which is covered by the macro BS.

assumptions:

- The system is an OFDM system with M' sub-channels, each of bandwidth b. These sub-channels are divided among the macro BSs based on conventional frequency reuse [16], i.e., given reuse factor r, the M' sub-channels are equally divided among the macro BSs such that each macro BS is granted a group of $M = \frac{M'}{r}$ sub-channels.
- The sub-channel gains are the same on each sub-channel for a given (location,BS) pair.
- Fast fading is averaged out at the link level, and the sub-channel gains account for the path loss and shadowing effects, i.e., the channel gains are random, but remain constant for a relatively long period of time.
- Each pico BS is connected to the macro BS via a high capacity wired backhaul.
- The maximum transmit powers of the macro (P_m) and pico (P_p) BSs are fixed and known a priori.
- Each user can associate with only one BS.
- All BSs are *active* at all time, i.e., there is no time at which a BS is not transmitting. Also, a BS uses all its available transmit power at all time.
- There are N fixed users in the system. These users are placed at random at N of the L possible user locations, i.e., we fix the realization[2]. All users' information,

[2]A realization corresponds to the random placement of the N users in the cell area based on a specific user distribution.

including the channel gains, are available so that the SINR to each user from each BS can be computed.

- The rate function $f(\cdot)$ for each BS is known so that given the SINRs, users' rates from all BSs can be computed. We do not make any restricting assumptions on f. We assume f is the same for each BS (though our framework does not depend at all on this assumption).

Sub-channels are the resources that we allocate to the different BSs, so that our global objective function is maximized. We study different resource allocation and interference management schemes, including three channel allocation strategies and one type of power allocation scheme well studied in the literature.

3.1.1 Channel Allocation

We study three channel allocation schemes as follows:

- *Co-channel deployment (*CCD*):* Each BS transmits on all the sub-channels.

- *Orthogonal deployment (*OD*):* K sub-channels are dedicated exclusively to the pool of pico BSs and $(M - K)$ sub-channels are dedicated to the macro BS. Each pico BS transmits on all the K sub-channels.

- *Partially shared deployment (*PSD*):* K sub-channels are shared by the macro and pico BSs and the other $(M - K)$ sub-channels are dedicated to the macro BS. Each pico BS transmits on all the K sub-channels.

3.1.2 Power Allocation

We assume that the total transmit powers of the macro (P_m) and pico (P_p) BSs are fixed and known. For *CCD* and *OD*, we assume that the power budget of a BS is shared equally among all channels allocated to this BS. For *PSD*, we assume that the macro BS uses the same transmit power budget P_p on the K channels shared with the pico BSs, and that it uses $(P_m - P_p)$ on the other $(M - K)$ sub-channels [51].

3.1.3 Physical Link Model

Let $\mathcal{N}$ denote the set of users in the system. The SINR of user $i \in \mathcal{N}$ from BS $j \in \mathcal{B} \cup \{0\}$ on each sub-channel (i.e., on the downlink) can be written as:

$$\gamma_{ij}^{(c)} = \frac{P_j^{(c)} \, G_{ij}}{N_0 + \sum_{h \in \mathcal{I}_j} P_h^{(c)} \, G_{ih}} \tag{3.1}$$

where $\mathcal{I}_j$ is the set of BSs transmitting on the same set of sub-channels (not including j) in the multi-tier system, $P_j^{(c)}$ is the transmit power of BS j on each of its sub-channels, N_0 is the additive white Gaussian noise power, and G_{ij} is the flat gain between user i and BS j that accounts for the path loss, shadow fading, antenna gain, and equipment losses. Note that given a reuse factor r and a RA scheme (i.e., CCD, OD, or PSD), $\mathcal{I}_j$ the set of BSs that use the same set of sub-channels as well as the transmit power of each BS on each sub-channel, $P_j^{(c)}$, can be determined. Then $\gamma_{ij}^{(c)}$ can be calculated for all $i \in \mathcal{N}$ and $j \in \mathcal{B} \cup \{0\}$.

As mentioned earlier, we assume there is a mapping function $f(\cdot)$ that maps the

SINR to the corresponding rate in bps (bit/second), i.e., $r_{ij}^{(c)} = f(\gamma_{ij}^{(c)})$ [3]. Next, we formulate our optimization problems, one for each of the three channel allocation schemes presented above.

3.2 Problem Formulations and Solution Techniques

As briefly mentioned earlier, we select proportional fairness as our global objective function, i.e., we maximize $\sum_i \log(\lambda_i)$ where λ_i is the throughput of user i. To compute λ_i, let r_{ij} denote user i's link rate from BS j (i.e., $r_{ij} = |\mathcal{C}_j| \times f(\gamma_{ij}^{(c)})$ where $\mathcal{C}_j$ is the set of (flat) sub-channels allocated to BS j) and let α_{ij} be the proportion of time that user i is scheduled on the downlink by BS j. We assume that a BS allocates all its sub-channels to a user at a given time (which is a reasonable assumption if the channels are flat). Let x_{ij} be equal to one if user i is associated with BS j, and let it be 0, otherwise. Hence, for all $i \in \mathcal{N}$, $\sum_{j \in \mathcal{B} \cup \{0\}} x_{ij}$ is equal to one. Note that we implicitly assume that each user i can hear at least one BS with a non-zero rate, i.e., there are no non-covered users in the system. Hence, λ_i is equal to $\sum_{j \in \mathcal{B} \cup \{0\}} (x_{ij}\alpha_{ij})\, r_{ij}$. Note that BS j allocates all its time among its associated users and hence, $\sum_{i \in \mathcal{N}} (x_{ij}\alpha_{ij}) = 1$.

We begin with the formulation for *co-channel deployment*. In this case, the problem is only one of optimal association and scheduling, i.e., the $\{x_{ij}\}$ and the $\{\alpha_{ij}\}$ are the only variables. The problem can be formulated as follows: given the *CCD* channel allocation, the M channels, the channel gains for the N fixed users, the rate function $f(\cdot)$, and the transmit powers, compute $\{x_{ij}\}$ and $\{\alpha_{ij}\}$ so as to maximize

[3] Note that the effect of channel bandwidth b is implicit in the rate function $f(\cdot)$.

the global proportional fairness objective:

$$\mathbf{P_{CCD}}: \quad \max_{\{x_{ij}\},\{\alpha_{ij}\}} \quad \sum_{i\in\mathcal{N}} \log(\lambda_i)$$

$$\text{subject to} \quad \lambda_i = \sum_{j\in\mathcal{B}\cup\{0\}} (x_{ij}\alpha_{ij})\, r_{ij} \ , \quad \forall i \in \mathcal{N} \tag{3.2a}$$

$$\sum_{j\in\mathcal{B}\cup\{0\}} x_{ij} = 1 \ , \quad \forall i \in \mathcal{N} \tag{3.2b}$$

$$\sum_{i\in\mathcal{N}} (x_{ij}\alpha_{ij}) = 1 \ , \quad \forall j \in \mathcal{B}\cup\{0\} \tag{3.2c}$$

$$0 \le \alpha_{ij} \le 1, \ x_{ij} \in \{0,1\} \ , \quad \forall i \in \mathcal{N}, \ \forall j \in \mathcal{B}\cup\{0\} \tag{3.2d}$$

where $r_{ij} = M \times f(\gamma_{ij}^{(c)})$, and $P_j^{(c)} = \frac{P_m}{M}$ if $j = 0$ and $P_j^{(c)} = \frac{P_p}{M}$ otherwise.

We assume that the backhaul infrastructure is not the bottleneck. More precisely, let C_j denote the capacity of the wired backhaul between pico BS j and the macro BS. For each feasible solution $\{x_{ij}\}$, we need to have $\sum_{i\in\mathcal{N}} x_{ij}\lambda_i \le C_j$ for all $j \in \mathcal{B}\cup\{0\}$. If C_j is sufficiently large, i.e., $\sum_{i\in\mathcal{N}} x_{ij}\lambda_i \ll C_j$ for all feasible solutions $\{x_{ij}\}$ to $\mathbf{P_{CCD}}$, these constraints will be satisfied, and they can be removed from the problem.

Before discussing this problem in more details, we formulate the problem of optimal user association and resource allocation for *orthogonal deployment* that allocates the first K sub-channels to the pico BSs and the rest to the macro BS. Each pico BS will be assigned K sub-channels. Note that $P_j^{(c)} = \frac{P_p}{K}$ for $j \in \mathcal{B}$, and $P_0^{(c)} = \frac{P_m}{M-K}$.

In the case of *OD*, we optimize the same objective function as for *CCD* with respect to the following variables: K, $\{x_{ij}\}$, and $\{\alpha_{ij}\}$. Note that the effect of K is implicit in $\gamma_{ij}^{(c)}$. The problem can be formulated as follows: given the *OD* channel allocation, the M channels, the channel gains for the N fixed users, the rate function

$f(\cdot)$, and the transmit powers, compute K, $\{\alpha_{ij}\}$, and $\{x_{ij}\}$ so as to maximize the proportional fairness objective:

$$\mathbf{P_{OD}}: \quad \max_{K,\{x_{ij}\},\{\alpha_{ij}\}} \quad \sum_{i\in\mathcal{N}} \log(\lambda_i)$$

$$\text{subject to} \qquad (3.2a) - (3.2d)$$

$$r_{i0} = (M-K) \times f(\gamma_{i0}^{(c)}),\ \forall i \in \mathcal{N} \tag{3.3a}$$

$$r_{ij} = K \times f(\gamma_{ij}^{(c)}),\ \forall j \in \mathcal{B},\ \forall i \in \mathcal{N} \tag{3.3b}$$

$$K \in \{1, \cdots, M-1\} \tag{3.3c}$$

For *PSD*, the macro BS transmitting on the K sub-channels can be considered as a new BS in the system. By doing this, we optimize the same objective function as for OD with respect to K, $\{x_{ij}\}$, and $\{\alpha_{ij}\}$, and we obtain a problem $\mathbf{P_{PSD}}$ similar to $\mathbf{P_{OD}}$. For brevity, we do not present the problem formulation for *PSD*.

Our objective is to solve these three problems exactly which is not going to be possible as we explain now. First, note that the proposed problem $\mathbf{P_{OD}}$ is a very complex problem. Some variables such as K are discrete while some others such as $\{\alpha_{ij}\}$ are continuous. Hence, it is hard to solve this problem as it is. Since $K \in \{1, \cdots, M-1\}$, a solution for $\mathbf{P_{OD}}$ can be obtained by solving $\mathbf{P_{OD}}$ iteratively for all possible values of K, and then selecting the best solution. In particular, let us define the optimal value of the objective function for $\mathbf{P_{OD}}$ for a given K, as $PF^{\star}(K)$. Hence, the solution for $\mathbf{P_{OD}}$ can be obtained by solving $\max_K \{PF^{\star}(K)\}$. Let $\mathbf{P'_{OD}}$ and $\mathbf{P'_{PSD}}$ be the problems obtained by fixing K. $\mathbf{P'_{OD}}$ (as well as $\mathbf{P'_{PSD}}$) reduces to a joint problem of optimal user association and scheduling (as is $\mathbf{P_{CCD}}$). These

three problems are non-convex integer programs and are NP-hard [52]. Hence, it is not possible to obtain exact solutions to these problems efficiently. Our goal is to transform these problems into convex problems for which the relaxed programs can be solved efficiently (i.e., for which upper bounds can be computed). To do so we tackle the problem in two steps. We will explain these steps for *OD*, but similar steps can be used for *PSD* and *CCD*. In the first step, we are going to show that $\mathbf{P'_{OD}}$ can be reduced to a pure optimal association problem by proving that for the optimal solutions, $x_{ij}\alpha_{ij}$ is equal to x_{ij}/N_j where $N_j = \sum_{i\in\mathcal{N}} x_{ij}$ is the number of users associated with BS j. This means that the global PF criteria yields a solution based on local PF at each BS (i.e., each BS offers the same amount of time to all its users). In the second step, we will transform this pure optimal association problem into a convex program whose solutions provide tight upper bounds on the solutions of $\mathbf{P'_{OD}}$.

STEP 1 : As mentioned earlier, we focus on OD although similar results hold for *CCD* and *PSD*. Assume each BS uses local PF scheduling. According to Lemma 1 [53], a BS assigns the same amount of time to its users.

Lemma 1 *[53] Let's assume there is one BS and all users have the same priority. Given resource allocation parameters including the number of sub-channels and the transmit power on each sub-channel, PF scheduling assigns equal proportion of time to all users.*

We can then formulate a new pure association problem called $\mathbf{P'^{\ell}_{OD}}$ as follows where

we assume that each BS schedules using local PF:

$$\mathbf{P'^{\ell}_{OD}}: \max_{\{x_{ij}\},\{N_j\}} \sum_{i\in\mathcal{N}} \log(\lambda_i)$$

subject to

$$\sum_{j\in\mathcal{B}\cup\{0\}} x_{ij} = 1 \ , \quad \forall i \in \mathcal{N} \tag{3.4a}$$

$$\lambda_i = \sum_{j\in\mathcal{B}\cup\{0\}} \left(\frac{x_{ij}}{N_j}\right) r_{ij} \ , \quad \forall i \in \mathcal{N} \tag{3.4b}$$

$$x_{ij} \in \{0,1\},\ N_j = \sum_{i\in\mathcal{N}} x_{ij} \ , \quad \forall j \in \mathcal{B}\cup\{0\}, \quad \forall i \in \mathcal{N}, \tag{3.4c}$$

where all r_{ij}'s can be computed beforehand and used as inputs to the optimization problem.

We say that two problems are equivalent if and only if an exact solution of one is an exact solution of the other.

Theorem 1 *Given $\mathcal{B}$, $\mathcal{N}$, M, the channel gains, the rate function, the parameters of the* OD*, i.e., K, $\boldsymbol{P'_{OD}}$ and $\boldsymbol{P'^{\ell}_{OD}}$ are equivalent.*

Proof : Please see Appendix.

Based on this theorem, we work now with $\mathbf{P'^{\ell}_{OD}}$. Note that $\mathbf{P^{\ell}_{CCD}}$ and $\mathbf{P^{\ell}_{PSD}}$ can be reduced to the *same* non-convex integer program since their differences are all summarized in the r_{ij}'s that can be computed beforehand.

STEP 2 : To obtain an upper bound for $\mathbf{P'^{\ell}_{OD}}$ we could try to simply relax the integrality constraints on $\{x_{ij}\}$ (i.e., we assume that $0 \leq x_{ij} \leq 1$ for all i, j) and try to solve the relaxed problem. However, even after relaxing the integrality constraints in $\mathbf{P'^{\ell}_{OD}}$, the problem remains non-convex. Note that non-convex programs cannot

be easily solved exactly. Fortunately, the structure of $\mathbf{P'^{\ell}_{OD}}$ is such that we can reformulate it into an integer convex problem as follows. Noting that all x_{ij}'s are binary variables and $\sum_{j\in\mathcal{B}\cup\{0\}} x_{ij} = 1$ for all users, there exists only one value of j, i.e. $\bar{j}$, for which $x_{i\bar{j}} = 1$ (i.e., $x_{ij} = 0$, $\forall j \neq \bar{j}$). Therefore, the objective function in $\mathbf{P'^{\ell}_{OD}}$ can be rewritten as follows:

$$\sum_{i\in\mathcal{N}} \log\left(\sum_{j\in\mathcal{B}\cup\{0\}} \frac{x_{ij}}{N_j}\ r_{ij}\right) = \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{B}\cup\{0\}} x_{ij}\,\log\left(\frac{r_{ij}}{N_j}\right). \tag{3.5}$$

Using this property, $\mathbf{P'^{\ell}_{OD}}$ can be reformulated into a convex integer program and the relaxed program (with respect to the integrality constraints on $\{x_{ij}\}$) can be solved efficiently even for large systems since it is a convex problem. Note that this problem is convex, and hence it can be solved to the desired precision in polynomial time [54]. This enables us to obtain upper bounds on the performance of $\mathbf{P^{\ell}_{CCD}}$, $\mathbf{P'^{\ell}_{OD}}$, and $\mathbf{P'^{\ell}_{PSD}}$ in terms of the global objective function, i.e., $\sum_i \log(\lambda_i)$, for large Hetnets that are composed of a large number of users, one macro BS, and many pico BSs. Although we are unable to show the tightness of these bounds analytically, we show numerically that $\mathbf{P'^{\ell}_{OD}}$ indeed provides a tight upper bound. We can verify the tightness of these upper bounds by finding a feasible solution for a given resource allocation and then comparing the corresponding performance metric $\sum_i \log(\lambda_i)$ for this feasible solution with the computed upper bound. To generate feasible solutions for a given RA, we will use simple association rules. It is important to note that the problems $\mathbf{P_{CCD}}$, $\mathbf{P_{OD}}$, and $\mathbf{P_{PSD}}$ can be used to provide the performance metric for a given association rule. Indeed, if the association rule is given, then the $\{x_{ij}\}$ is given and the problems can then be solved easily. We will use this fact to compare

the performance of several simple user association rules under our three resource allocation schemes, as shown in Section 3.4.

The purpose of the static study is threefold: First, we want to compare the three resource allocation schemes, i.e., *CCD*, *PSD*, and *OD* not only in terms of the objective function, but also in terms of aggregate throughput, and minimum throughput in the system, i.e., the performance metrics are $\sum_i \log(\lambda_i)$, $\sum_i \lambda_i$, and $\min_i \{\lambda_i\}$. Operators are typically trying to trade-off fairness (usually using proportional fairness criteria), the total aggregate throughput which is a measure of the "capacity" of their system, and some criteria to take edge users' performance into consideration. We chose to use the minimum rate in the system as such a measure. Second, we want to study how different simple association rules perform as compared to the optimal solutions for these three resource allocation schemes. Finally, we want to study in more details the impact of some of the parameters of *PSD* which, under our assumptions, performs significantly better than *CCD* and *OD*. Next, we describe the simple association rules that we are going to study and compare.

3.3 Simple User Association Rules

In practical cellular systems, users arrive in the network, stay for a while, and depart the network. Such systems would work optimally if we are able to compute the optimal RA parameter (if any) and associate and re-associate users optimally whenever a new user arrives, a user moves or departs the system, or the channel gains change significantly. Such heavy computations are difficult to do online and changing the RA parameter and re-associating a large number of users frequently might degrade

the system's performance and result in oscillations. To avoid such possible problems, simple association rules have been used in homogeneous cellular systems and proposed in the literature for heterogeneous systems. These rules typically associate users based on physical layer parameters without considering other system's issues such as load balancing among BSs. We study some of those rules and propose a new user association rule that we call *Picocell First*. A description of these rules is as follows:

1. **Best SINR**: A user i associates with BS $j^\star$ that provides the highest SINR, i.e., $j^\star = \arg\max_{j\in\mathcal{B}\cup\{0\}} \{\gamma_{ij}^{(c)}\}$ where $\gamma_{ij}^{(c)}$ denotes the SINR of user i from BS j, on each sub-channel respectively. This association rule has been used in conventional cellular networks.

2. **Range Extension (RE) [55]:** A user at location i associates with BS $j^\star = \arg\min_{j\in\mathcal{B}\cup\{0\}} \{\delta_{ij}\}$ where δ_{ij} is the path loss from BS j to location i.

3. **Picocell First (PCF)** [50]: A user at location i associates with pico BS $j^\star = \arg\max_{j\in\mathcal{B}\cup\{0\}} \{\gamma_{ij}^{(c)}\}$ as long as $\gamma_{ij^\star} > \beta$ where β is a tuning parameter. Note that $\gamma_{ij}^{(c)}$ denotes user i's SINR on each sub-channel. If $\max_{j\in\mathcal{B}} \{\gamma_{ij}^{(c)}\} < \beta$, user i associates with the BS that provides the highest SINR. This rule associates users with pico BSs regardless of their received power from the macro BS as long as the best SINR seen from a pico BS is larger than β. The motivation behind this rule is to bring BSs closer to users and offload data traffic via pico BSs.

For each of these rules, once the physical layer parameters are known, we can compute the values of x_{ij} for all users i and BSs j. To compute the physical layer parameters, we need to fix the resource allocation scheme and its parameters if any,

i.e., K, for *OD* and *PSD*. Therefore, for *OD* and *PSD*, to compute the system's performance when the user association is given by a simple association rule, we need to fix K, and to compute the system's performance corresponding to these parameters, and then iterate on these parameters. Note that for *CCD* the resource allocation parameters are fixed. Thus, given a user association $\{x_{ij}\}$, we can compute the solution to $\mathbf{P}^{\ell}_{\mathbf{CCD}}$ by calculating $\sum_i \log(\lambda_i)$.

We now have a unified framework, i.e., the proposed joint user association and resource allocation problems, and we can compute upper bounds on the objective function of the proposed problems. Using this framework, we can compute the optimal resource allocation parameters and the performance metrics (i.e., $\sum_i \log(\lambda_i)$, $\sum_i \lambda_i$, $\min_i \{\lambda_i\}$) when an association rule is given. Note that when we fix the association rule, we can generate a feasible integral solution to each problem $\mathbf{P_{CCD}}$, $\mathbf{P_{PSD}}$, and $\mathbf{P_{OD}}$. If we can find a simple association that yields an objective function close to the corresponding upper bound, then we would have validated the tightness of our bound and the goodness of that simple association rule. Next, we explore the performance of existing and proposed user association and RA schemes.

3.4 Numerical Results

3.4.1 Parameter Settings

We consider a communication system composed of 19 macro BSs. Each macro BS's coverage area is overlaid with four pico BSs. The network has an inter-cell distance of 500 m. We study the cell at the center which is a hexagonal area with radius

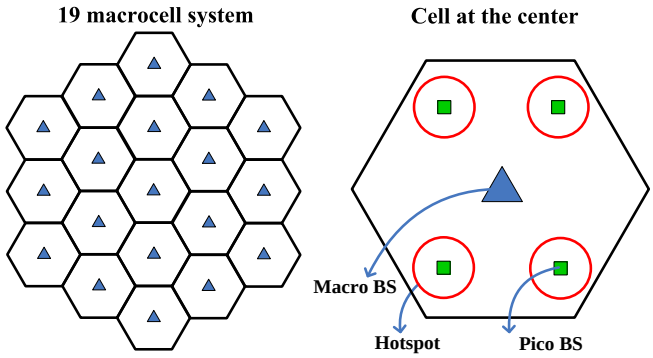


Figure 3.1: A Hetnet comprising 19 macro BSs (the triangles). Each macro BS is overlaid with 4 pico BSs. Pico locations for the cell at the center are shown in the figure in the right-hand side. The triangle is the macro BS and the squares are the pico BSs. There are four hotspots shown as circles around the pico BSs.

$R = 500/\sqrt{3}$ m (see Fig. 3.1). The cell is covered by one macro BS and four pico BSs. The macro BS is located at the center of the cell while the pico BSs are located symmetrically around the macro BS with distance $d = 230$ m from the center. As mentioned earlier, we assume that the system is an OFDM system with M' sub-channels. We consider a reuse factor of "three", i.e., each macro BS has access to $M = M'/3$ sub-channels. We use $M = 100$ sub-channels, and assume that there are $L = 2000$ possible user locations in each macro BS area.

We consider two configurations. In Configuration 1, there are $N = 20$ users while in Configuration 2, there are $N = 60$ users distributed in the cell area. To consider the case where users are clustered in some areas in the system, we consider two types of user distribution: uniform (UD) and non-uniform (NUD). With the uniform user distribution, the N users are uniformly distributed at random in the cell area, while with the non-uniform user distribution, $\frac{2}{3}$ of the users are uniformly distributed at random in the hotspot areas shown in Fig. 3.1 while the rest is distributed uniformly

Table 3.1: Physical Layer Parameters

Noise Power	$-174\ \frac{\text{dBm}}{\text{Hz}}$	T_{subframe}	1 ms
P_{pico}	30 dBm	P_{macro}	46 dBm
UE Ant. Gain	0 dB	Sub-channel Bandwidth	180 KHz
Shadowing s.d.	8 dB	Penetration Loss	20 dB
SC_{ofdm}	12	SY_{ofdm}	14
Path Loss (Pico)	$140.7 + 36.7 \log_{10}(d/1000),\ d \geq 10m$		
Path Loss (Macro)	$128 + 37.6 \log_{10}(d/1000),\ d \geq 35m$		

Table 3.2: Modulation and Coding Schemes-LTE

SINR thresholds (in dB)	-6.5	-4	-2.6	-1	1	3	6.6	10	11.4	11.8	13	13.8	15.6	16.8	17.6
Efficiency (in bits/symbol)	0.15	0.23	0.38	0.60	0.88	1.18	1.48	1.91	2.41	2.73	3.32	3.90	4.52	5.12	5.55

in the cell area. Each hotspot is a circle of radius 80m, centered at a pico BS.

The physical layer parameters are based on the 3GPP evaluation methodology document [56] used for Hetnets in LTE. These parameters are shown in Table 3.1. We use the SINR model introduced in Section 3.1.3 that accounts for path loss and slow fading. Slow fading is modeled by a log-normal shadowing with standard deviation of 8 dB, and path losses for pico and macro BSs are given in Table 3.1. We assume that the system uses adaptive modulation and coding with 15 discrete rates. Table 3.2 taken from [57] and [58] gives us the mapping between the SINR and the efficiency (in bits/symbol) per sub-carrier for the modulation and coding schemes (MCS) for LTE. The bit rate obtained by a user that has a SINR between level ℓ and level $\ell + 1$ is $r = \frac{\text{SC}_{\text{ofdm}}\ \text{SY}_{\text{ofdm}}}{T_{\text{subframe}}} e_\ell$ where e_ℓ is the efficiency (bits/symbol) of the corresponding level ℓ, SC_{ofdm} is the number of data subcarriers per sub-channel bandwidth, SY_{ofdm} is the number of OFDM symbols per subframe, and T_{subframe} is the subframe duration in time units. The value of these parameters are shown in Table 3.1.

Table 3.3: The SINR threshold values for the tuning parameter β

β	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
SINR thresholds (in dB)	-6.5	-4	-2.6	-1	1	3	6.6	8.17	9.33	10.24

Table 3.4: The SINR threshold values for the tuning parameter β

β	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
SINR thresholds (in dB)	10.99	11.63	12.19	12.68	13.13	13.53	13.8	15.6	16.8	17.6

Our comparisons are based on the following performance metrics:

- GM : Geometric mean rate of the users, i.e., $\sqrt[N]{\prod_{i=1}^{N} \lambda_i}$ (note that for fixed N, maximizing the GM is equivalent to maximizing our objective function);
- Min Throughput : Minimum throughput among all users, i.e., $\min_i\{\lambda_i\}$;
- TT : Total throughput of the system, i.e., $\sum_{i=1}^{N} \lambda_i$.

"Picocell First" has a tuning parameter β . We assume that β can take any one of the SINR threshold values shown in tables 3.3-3.4. In the numerical results, we select the value of β that gives the highest possible geometric mean rate. Two configurations and two types of user distribution provide us four scenarios to compare the performance of different combinations of resource allocation and user association schemes. For each scenario, we compute the upper bound for 20 networks. A network corresponds to the random realization of the shadowing coefficients for the $L = 2000$ locations from all the BSs in the multi-tier system. For each network, we compute the average results over 100 realizations. A realization corresponds to the random

placement of the N users in the system area based on the scenario. In the following, we show the trends averaged over the 20 networks. For each network, we compute the average results over the 100 realizations.

3.4.2 Comparison Results

Tables 3.5-3.6 provide the results for four typical scenarios corresponding to uniform and non-uniform user distributions, respectively. In the row entitled "GM relaxation" in these tables, the upper bounds of the joint user association and resource allocation for *CCD*, *OD*, and *PSD* are provided when all system parameters are computed optimally (i.e, for *OD* and *PSD*, we compute the best K). To check the tightness of these upper bounds, we compare them to the geometric mean rate of the Picocell First association rule (*GM PCF* in the tables) computed for the β that provides the highest geometric mean rate. For the scenarios, we also consider the system without any pico BS (called "No pico" in tables 3.5-3.6) to see how much gain can be achieved by deploying pico BSs. The results show the following:

- *PSD* and *OD* perform significantly better than *CCD* in all cases. *PSD* and *OD* perform almost the same with a slight advantage for *PSD*. For *PSD*, we saw gains (with respect to *CCD*) in total throughput in the range of 50% to 110%, and gains in geometric mean rate in the range of 35% to 88% over 20 networks. This is not so surprising under our assumption that the BSs are not coordinated and transmit at all time on all channels allocated to them at full power.

- The association rule *Picocell First* is almost optimal since the geometric mean rate of the Picocell First is very close to the upper bound for *CCD*, *OD*, and

for *PSD* when all system parameters, including β, are chosen optimally. This has two consequences. It validates our relaxation approach because an integer feasible solution to the proposed problems achieves almost the same geometric mean rate as the solution of the relaxed problem. It also shows that *Picocell First* is a very good yet simple association rule. Similar results were obtained in [50] for a different framework. Since it is near optimal, we will use *Picocell First* when we want to compare the resource allocation schemes in term of minimum throughput and total throughput.

- The comparison of the system's performance (using *Picocell First*) between the system with and without pico BSs ("No pico" in the tables) shows that pico BSs can significantly improve the performance of the system. We saw gains (with respect to the system without pico BSs) in total throughput in the range of 250% to 340%, and gains in geometric mean rate in the range of 100% to 250% over 20 networks.

3.4.3 In Depth Study of PSD

We now study *Partially Shared deployment* in more details. We compare the performance of the simple association rules with the upper bound as a function of K. For each value of K, we compute the upper bound, i.e., the solution to the relaxed problem $\mathbf{P'}^{\ell}_{\mathbf{PSD}}$, and the corresponding geometric mean rate for each association rule. The results for the four scenarios are shown in figures 3.2 to 3.5, which all show the same relative performances. The curve corresponding to the upper bound is labeled *Upper bound* in the figures. Since *CCD* is often considered as the preferred option

Table 3.5: Configuration 1: The average GM, Min Throughput, and TT, in bits per second for $N = 20$, $M = 100$, and best β. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

Uniform user distribution

RA scheme	CCD	OD	PSD
GM relaxation	3.6867e+6	5.2236e+6	5.8960e+6
GM PCF	3.6846e+6	4.6088e+6	5.0266e+6
GM No Pico	2.6299e+6	2.6299e+6	2.6299e+6
Min Throughput PCF	0.9105e+6	1.2018e+6	1.1678+6
Min Throughput No Pico	0.6923e+6	0.6923e+6	0.6923e+6
TT PCF	3.5473e+8	5.7253e+8	6.4054e+8
TT No Pico	1.8765e+8	1.8765e+8	1.8765e+8

Non-uniform user distribution

RA scheme	CCD	OD	PSD
GM relaxation	4.0656e+6	6.6021e+6	7.2058e+6
GM PCF	4.0221e+6	6.0025e+6	6.5796e+6
GM No Pico	2.3907e+6	2.3907e+6	2.3907e+6
Min Throughput PCF	0.9270e+6	1.3662e+6	1.5721e+6
Min Throughput No Pico	0.6325e+6	0.6325e+6	0.6325e+6
TT PCF	3.9391e+8	6.5226e+8	7.6344e+8
TT No Pico	1.7407e+8	1.7407e+8	1.7407e+8

Table 3.6: Configuration 2: The average GM, Min Throughput, and TT, in bits per second for $N = 60$, $M = 100$, and best β. Note that the results are averaged over 20 networks, and that for each network, we compute teh average results over 100 realizations.

Uniform user distribution

RA scheme	CCD	OD	PSD
GM relaxation	1.2876e+6	1.8365e+6	2.0497e+6
GM PCF	1.2875e+6	1.6516e+6	1.8808e+6
GM No Pico	0.8695e+6	0.8695e+6	0.8695e+6
Min Throughput PCF	0.2023e+6	0.2909e+6	0.3557+6
Min Throughput No Pico	0.1467e+6	0.1467e+6	0.1467e+6
TT PCF	1.4272e+8	2.2405e+8	2.3142e+8
TT No Pico	0.6259e+8	0.6259e+8	0.6259e+8

Non-uniform user distribution

RA scheme	CCD	OD	PSD
GM relaxation	1.4394e+6	2.3364e+6	2.5779e+6
GM PCF	1.4308e+6	2.1973e+6	2.3778e+6
GM No Pico	0.7742e+6	0.7742e+6	0.7742e+6
Min Throughput PCF	0.2271e+6	0.4291e+6	0.4095e+6
Min Throughput No Pico	0.1293e+6	0.1293e+6	0.1293e+6
TT PCF	1.4810e+8	2.3139e+8	2.6410e+8
TT No Pico	0.5713e+8	0.5713e+8	0.5713e+8

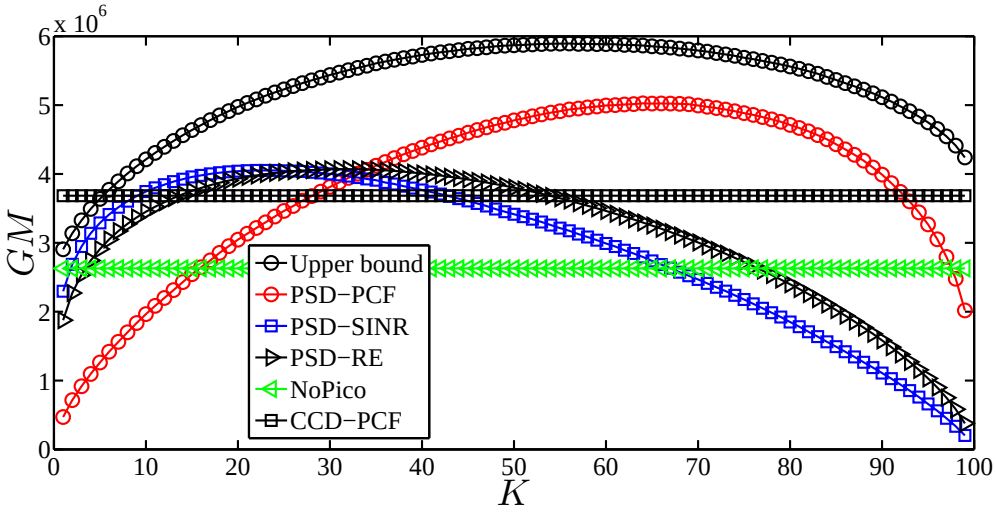


Figure 3.2: Configuration 1, PSD, UD: Geometric mean rate (in bits per second) as a function of K when $N = 20$ and $M = 100$. We choose the β that provides the highest geometric mean rate. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

in Hetnets, we also show the upper bound for *CCD* for each scenario to compare its performance with *PSD*. Note that Picocell First performs very well for CCD, i.e., it is very close to the upper bound (see tables 3.5-3.6). Because of this, we show the performance of Picocell First for *CCD* instead of the upper bound for *CCD*. The comparison between the upper bounds for *PSD* and *CCD* for the four scenarios, shows that *PSD* performs better than *CCD* for a large range of K, i.e., even if the operator cannot choose K optimally, he should still prefer *PSD* over *CCD* under our assumptions. Figures 3.2 to 3.5 also show that the optimal value of the channel allocation parameter K is highly dependent on the deployed association and on the scenario at hand.

The comparison between the geometric mean rate of the simple association rules

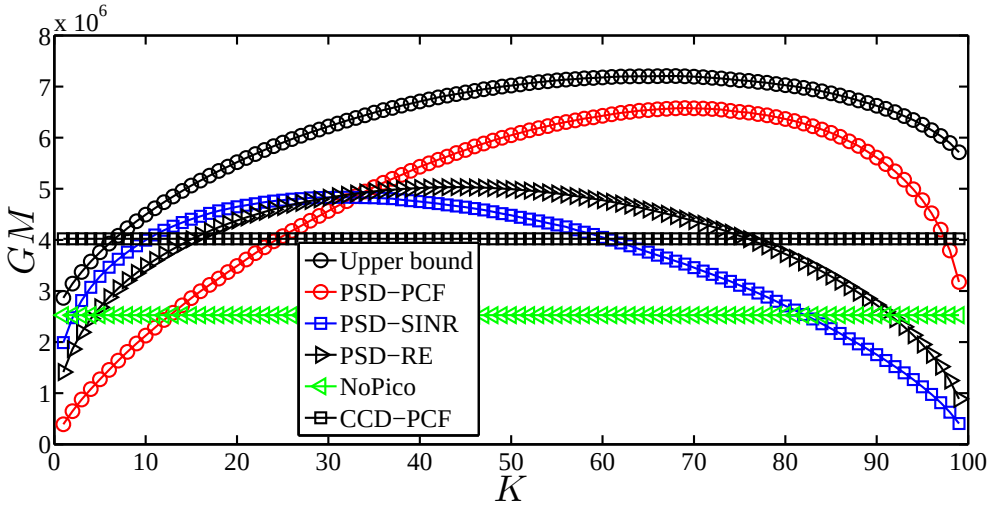


Figure 3.3: Configuration 1, PSD, NUD: Geometric mean rate (in bits per second) as a function of K when $N = 20$ and $M = 100$. We choose the β that provides the highest geometric mean rate. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

and the upper bound for *PSD* shows that "Picocell First" almost always performs the best of the three rules for a range of values of K. The results also show that "Best SINR" does not perform well in any of the scenarios. When K is not chosen optimally, the performance of "Picocell First" can be far from the upper bound. We believe that this can be explained by the fact that if resource allocation is not performed well, load balancing becomes a major issue and none of our simple association rules take load balancing into account.

3.5 Conclusions

We have studied the problem of joint user association and resource allocation in Hetnets that consist of macro and pico BSs. We have considered three channel allocation

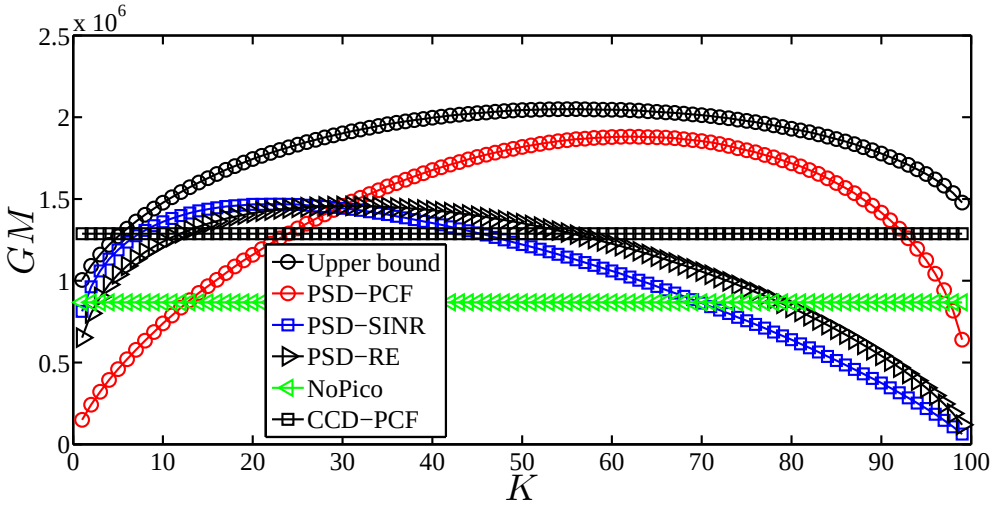


Figure 3.4: Configuration 2, PSD, UD: Geometric mean rate (in bits per second) as a function of K when $N = 60$ and $M = 100$. We choose the β that provides the highest geometric mean rate. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

schemes, and assumed that all the BSs are transmitting all the time on all their allocated channels. The proposed problems are non-linear integer programs, and hence it is not possible to efficiently obtain exact solutions. Therefore, we have developed techniques to obtain upper bounds on the system's performance. Numerical results show that the proposed upper bounds are tight and can be used as benchmarks to quantify how well different user association rules and resource allocation schemes perform.

Our numerical results indicate that significant performance gains are achievable for Hetnets with a static user population if the system uses the right combination of user association and resource allocation. Gains in total throughput in the range of 250% to 340%, and gains in geometric mean rate in the range of 100% to 250%, are

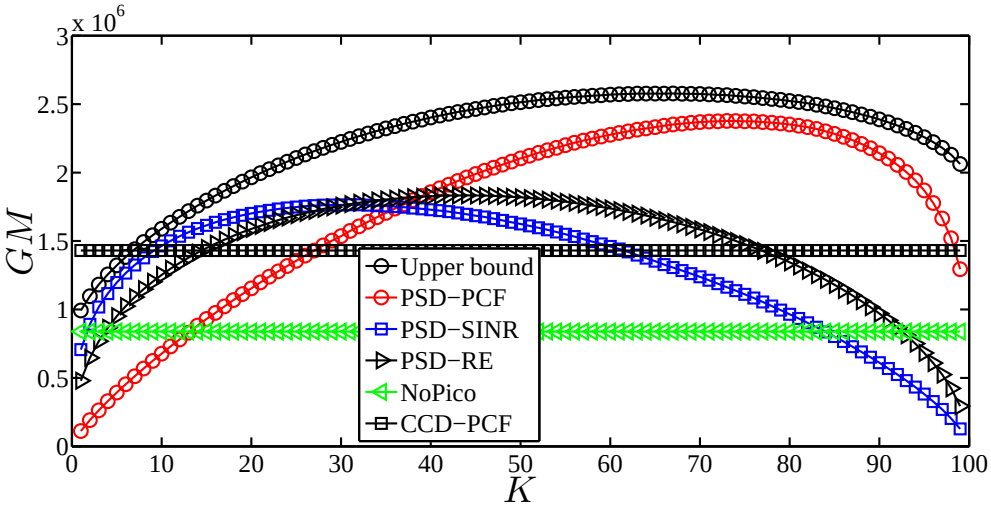


Figure 3.5: Configuration 2, PSD, NUD: Geometric mean rate (in bits per second) as a function of K when $N = 60$ and $M = 100$. We choose the β that provides the highest geometric mean rate. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

achievable for Hetnets using pico BSs. *Partially shared deployment* and *orthogonal deployment* perform significantly better than *co-channel deployment*. Noting the significant impact of association rules on the performance of Hetnets, we have proposed a new user association rule. Our results show that rules which favor associating users with pico BSs (e.g. "Picocell First" and "Range Extension") yield *significantly better* performance than the conventional association rule if their tuning parameters are chosen properly and if the resource allocation parameters have been chosen optimally.

Chapter 4

Hetnets with a Dynamic User Population

4.1 System Model

We consider a multi-tier communication system composed of several macro BSs. Each macro BS is overlaid with B pico BSs. We focus on the cell at the center (cell 0), and assume that the cell is overlaid with B pico BSs that are identical in terms of transmit power, antenna gain, and backhaul capacity. The coverage area of the macro BS is covered by a grid corresponding to possible locations for the users (i.e., we discretize the set of locations at which users can be). Let $\mathcal{B}$ and $\mathcal{L}$ denote the set of pico BSs deployed by the operator ($|\mathcal{B}| = B$) and the set of possible user locations within the cell area ($|\mathcal{L}| = L$), respectively. We focus on the downlink, and make the following assumptions:

1. The system is an OFDM system with M' sub-channels, each of bandwidth b.

These sub-channels are divided among the macro BSs based on conventional frequency reuse [16], i.e., given reuse factor r, the M' sub-channels are equally divided among the macro BSs such that each macro BS is granted a group of $M = M'/r$ sub-channels.

2. The sub-channel gains are the same on each sub-channel for a given (location,BS) pair.

3. Fast fading is averaged out at the link level, and the sub-channel gains account for the path loss and shadowing effects, i.e., the channel gains are random, but remain constant for a relatively long period of time.

4. Each pico BS is connected to the macro BS via a high capacity wired backhaul.

5. The maximum transmit powers of the macro (P_m) and pico (P_p) BSs are fixed and known a priori.

6. The total transmit power of each BS is shared equally among all the sub-channels allocated to it.

7. Each BS schedules its users using local PF, and there is no time at which the BS is not transmitting (i.e., all BSs are *active* all the time).

8. Each user associates with one BS.

9. Users arrive at each location $i \in \mathcal{L}$ according to a Poisson process with density $\bar{\lambda}$. Users arrival process into the entire cell area is a Poisson process with density $\lambda = |\mathcal{L}|\bar{\lambda}$.

10. Users do not move during their calls, and hence the link rate between a user and its BS is fixed.

11. Users arriving to the system download files whose sizes are independent and identically distributed (i.i.d.) random variables of mean F. Note that the file size has a general distribution and is not necessarily exponentially distributed.

12. Users depart the system as soon as their files have been downloaded completely.

Sub-channels are the resources that we allocate to different BSs, so that the network performance metrics are optimized. We consider three different resource allocation and interference management schemes:

- *Co-channel deployment (*CCD*):* Each BS transmits on all the M sub-channels.

- *Orthogonal deployment (*OD*):* K sub-channels are dedicated exclusively to the pool of pico BSs and the remaining $(M - K)$ sub-channels are dedicated to the macro BS. Each pico BS transmits on all the K sub-channels, and the macro BS transmits on all the $(M - K)$ sub-channels.

- *Partially shared deployment (*PSD*):* K sub-channels are shared by the macro and pico BSs and the other $(M - K)$ sub-channels are dedicated to the macro BS. Each pico BS transmits on all the K sub-channels, and the macro BS transmits on all the M sub-channels.

For *CCD* and *OD*, we assume that the power budget of a BS is shared equally among all channels allocated to this BS. For *PSD*, we assume that the macro BS uses the same transmit power budget P_p on the K channels shared with the pico BSs, and that it uses $(P_m - P_p)$ on the other $(M - K)$ sub-channels.

The deployed channel allocation strategy determines the set of BSs transmitting on the same set of sub-channels as well as the transmit power of each BS on each of its sub-channels. Knowing the set of co-channel BSs and the transmit power of each BS on each of its sub-channels, we can compute the SINR of location $i \in \mathcal{L}$ from BS $j \in \mathcal{B} \cup \{0\}$ on each sub-channel (call it $\gamma_{ij}^{(c)}$) as follows:

$$\gamma_{ij}^{(c)} = \frac{P_j^{(c)} \; G_{ij}}{N_0 + \sum_{h \in \mathcal{I}_j} P_h^{(c)} \; G_{ih}} \tag{4.1}$$

where $\mathcal{I}_j$ is the set of BSs transmitting on the same set of sub-channels (not including j) in the multi-tier system, $P_j^{(c)}$ is the transmit power of BS j on each of its sub-channels, N_0 is the additive white Gaussian noise power on the sub-channel, and G_{ij} is the flat gain between location i and BS j that accounts for the path loss, shadow fading, antenna gain, and equipment losses. Note that given a reuse factor r, and a resource allocation and its parameter if any, i.e., K for OD or PSD, $\mathcal{I}_j$ (the set of BSs that use the same set of sub-channels) as well as $P_j^{(c)}$ (the transmit power of each BS on each sub-channel) can be determined. Given $\gamma_{ij}^{(c)}$ for $i \in \mathcal{L}$ and $j \in \mathcal{B} \cup \{0\}$, using the mapping function $f(\cdot)$ that maps the SINR to the corresponding link rate in bps (bit/second), the link rate at location i from BS j on each sub-channel can be calculated by $r_{ij}^{(c)} = f(\gamma_{ij}^{(c)})$[1].

A decision to associate a user with a BS will affect the throughput seen by the user as well as the throughput of the other users associated with that BS. Each location in the cell area, might be in the coverage area of multiple BSs, and hence the users arriving to a location can possibly associate with different BSs. An user

[1]Note that the effect of channel bandwidth b is implicit in the rate function $f(\cdot)$.

association policy UA defines a set of rules for assigning users to the different available BSs in the Hetnet. Several user association rules have been proposed in the literature (e.g., [46], [48], [50], [59]- [63]) that perform well for a static user population. Typically, such rules use physical layer parameters (averaged at the link level) such as received signal power, path loss, and SINR, to choose the best BS for each user. Since fast fading is averaged out at the link level, the physical layer parameters comprise only the path loss and shadowing effects which are quasi-static. Therefore, the users arriving at a certain location, would associate with the same BS if the system associates users based on their physical layer parameters. In this study, for simplicity reasons, we assume that the users arriving at a certain location, associate with the same BS. Under this assumption, an user association policy defines a set of rules for assigning each location to a unique BS in the Hetnet.

Users arrive at each of the L possible locations according to a Poisson process with density $\bar{\lambda}$ (i.e., the inter-arrival times at each location are exponentially distributed random variables), and download files. Using a certain UA rule, the arriving users associate with one of the BSs in the cell area, and get served by the BS. Each BS offers the same amount of time to all its users since it uses local PF scheduling and its backhaul capacity is infinite. Therefore, the arriving users to each BS get served based on the processor sharing (PS) discipline [64].

The processor sharing discipline, introduced in [64], captures the following properties of time-sharing systems. First, it allows multiple users to be in service at the same time so that the server shares its resources equally among the users present in the system. Second, it allows all arriving users to enter service immediately. Therefore, in the systems using the GPS discipline, the service rate allocated to each user

depends on the total number of users present in the system. For more information on the processor sharing discipline, we refer the reader to [65]- [67].

We can capture the system dynamics by a queueing model which takes into account the users' arrival and departure processes as well as the scheduling policy. We consider the coverage area of each BS as a generalized processor sharing queue, and each of the L locations in the cell area as a class. We assume that each location sees a fixed SINR that can be computed using (4.1). Therefore, each location i (i.e., class i) has its own general service time distribution which is dependent on the file size distribution in location i. More precisely, we model the coverage area of each BS as a GPS queue, and the Hetnet as a set of GPS queues serving the cell area. Since each BS transmits all the time and there is no coordination among the BSs, we can decompose the set of GPS queues, and view the Hetnet as a set of independent GPS queues serving the cell area.

Users arrive in the system according to a Poisson process and download files whose sizes are i.i.d. random variables of mean F. Therefore, each BS can be modeled by a multi-class M/G/1 PS queue. The stationary distribution of the number of users in the M/G/1 PS queue has a geometric distribution and is insensitive to the service time distribution except through its mean. Using this property, we can assume that a user arriving to location i requires to download a file of data with an exponentially distributed file size of mean F, and hence we can view each BS j as a multi-class M/M/1 PS queue. Using this model, the Hetnet can be viewed as a set of multi-class M/M/1 PS queues serving the cell area.

In each of the multi-class M/M/1 PS queues, the per user service rate is a function of the number of channels available at the BS, the level of interference, and the current

amount of users associated with the BS. Therefore, the deployed UA and RA schemes will have a critical impact on service rates (as well as other delay-based performance metrics), and this is what we study in this chapter. Our performance metrics are as follows:

1. *The Highest Possible Arrival Rate $\bar{\lambda}_{max}$:* When user arrivals happen faster than service completions (i.e., file downloads) in a queue, the queue length can grow indefinitely long (i.e., the queue will not have a stationary distribution). As mentioned above, service rates are mainly determined by the deployed RA and UA. Given a set of service rates, arrival rate $\bar{\lambda}$ should be less than or equal to a certain value (call it $\bar{\lambda}_{max}$)[2] to make sure that the system is stable. The value of $\bar{\lambda}_{max}$ is highly dependent on the service rates. The per-user service rate is mainly determined by the number of channels available at the BS, the level of interference, and the current amount of users associated with the BS which are a function of the deployed UA and RA. Therefore, the network processes (i.e., UA and RA) will have a critical impact on $\bar{\lambda}_{max}$, and this is what we want to study.

2. *Network Delay:* The average delay experienced by the users arriving to location i (we call it the average delay of class i) depends on the service rate in location i as well as the arrival rate $\bar{\lambda}$. As mentioned earlier, service rates are mainly determined by the deployed RA and UA. Therefore, given an arrival rate $\bar{\lambda}$, the average delay of class i is mainly determined by the deployed UA and RA. We choose the maximum average delay per class and the average delay in the cell

[2]Let λ_{max} denote the highest arrival rate into the entire cell area.

as our delay metrics, and study the interplay of UA, RA, and delay when the arrival rate (i.e., $\bar{\lambda} \leq \bar{\lambda}_{max}$) is given. The maximum average delay per class can be seen as a performance metric for edge users (i.e., users with low link rates) while the average delay in the cell represents the average delay over all users arriving in the cell.

3. *Delay-Constrained Maximum Throughput $\bar{\lambda}_T$:* Arrival rate $\bar{\lambda}$ can be as large as $\bar{\lambda}_{max}$ if there is no constraint on delay. When there is a constraint on delay (e.g., the maximum average delay per class should be less than or equal to T), the operator has to control the arrival process (i.e., the operator needs to perform an admission control) since the average delay is increasing in $\bar{\lambda}$. We select delay-constrained maximum throughput $\bar{\lambda}_T$ as our third metric, and try to understand the impact of UA and RA on this metric.

These performance metrics are highly dependent on the users' service rates as well as the user arrival rate. As mentioned earlier, the level of interference and the quantity of channels available at each BS, and the number of users associated with each BS have a critical impact on the service rates. Therefore, the deployed UA and RA will significantly impact the values of our performance metrics.

We define the Hetnet that we study in terms of UA, RA, user arrival, and our performance metrics, and try to answer the following questions:

- What is the highest possible arrival rate $\bar{\lambda}_{max}$ for a given resource allocation and its parameter if any? How is this metric impacted by the deployed UA?

- In the case, where admission control cannot be performed, the operator has to choose schemes that yield good delay performance over a wide range of $\bar{\lambda}$. What

is the best combination of UA and RA schemes to obtain the lowest possible delay? How are these delay metrics affected by different UA and RA schemes?

- Assuming that the operator could perform admission control, how is $\bar{\lambda}_T$ a function of T for different UA and RA?

These questions represent different possible scenarios that a network operator might need to consider during the engineering of its Hetnet. To answer these questions, we fix the resource allocation and its parameter if any (i.e., either *CCD* or, *OD* or *PSD* with a given K), and compute the service rates. Given the service rates, we try to answer each of the above questions. By answering the above questions iteratively for all possible values of K, we can answer the above questions completely, i.e., we can find the best resource allocation and the value of its parameter if any.

4.2 Problem Formulations

We focus on the long-run performance of the Hetnet, i.e., the set of M/M/1 PS queues. The long-run performance of a single multi-class M/G/1 PS queue is very well studied in the literature. In [68], Karray *et al.* use the fact that the stationary distribution of the number of users in an M/G/1 PS queue is insensitive to the service time distribution except through its mean, and derive the distribution of the number of users in a cell (i.e., M/M/1 PS queue) as well as the expected delay per location in the cell. We use the queueing results in [68] without proof.

4.2.1 Network Stability

Let us fix the resource allocation and its parameter if any, i.e., K for OD or PSD. Let $\boldsymbol{RA}$ denote the fixed resource allocation. Given the resource allocation (i.e., $\boldsymbol{RA}$), our objective is to compute the highest possible arrival rate $\overline{\lambda}_{max}(\boldsymbol{RA})$. As mentioned earlier, the highest possible arrival rate is linked to stability, and the Hetnet can be viewed as a set of independent M/M/1 PS queues. Therefore, the system is stable *if and only if* each queue[3] is stable. Before presenting the stability criterion for an M/M/1 PS queue, let us define the load factor of BS j (call it ρ_j). To do this, let r_{ij} denote location i's link rate from BS j (i.e., $r_{ij} = |\mathcal{K}_j| f(\gamma_{ij}^{(c)})$ where $\mathcal{K}_j$ denotes the set of sub-channels allocated to BS j). Let $x_{ij} = 1$ if location i is associated with BS j, and let it be 0, otherwise. Hence, for all $i \in \mathcal{L}$, $\sum_{j \in \mathcal{B} \cup \{0\}} x_{ij} = 1$. We define the load factor of BS j as follows:

$$\rho_j = \sum_{i \in \mathcal{L}} x_{ij} \frac{\bar{\lambda} F}{r_{ij}} \quad j \in \mathcal{B} \cup \{0\} \ . \tag{4.2}$$

An M/M/1 PS queue is stable *if and only if* the load factor of the queue is strictly less than one [68]. Using the stability criterion, the Hetnet is stable *if and only if* the following condition is satisfied:

$$\bar{\lambda} \sum_{i \in \mathcal{L}} x_{ij} \frac{F}{r_{ij}} < 1 \ , \quad \forall j \in \mathcal{B} \cup \{0\}. \tag{4.3}$$

Given the resource allocation, i.e., either CCD, or OD or PSD with a given K, we can compute $\{r_{ij}\}$. Therefore, given the resource allocation, an arrival rate $\bar{\lambda}$ is feasible

[3]We will use the term queue and BS interchangeably in the thesis.

(i.e., the system is stable) *if and only if* there exists an user association $\{x_{ij}\}$ for which (4.3) is satisfied. By using (4.3), we can easily check the feasibility of $\bar{\lambda}$ for a given user association, but it is harder to find whether there exists an user association for which (4.3) is satisfied. In order to write tractable optimization problems, we prefer not to work with strict inequalities[4], and hence we introduce a parameter $\bar{\rho}$, and assume that the load at each BS j cannot be larger than $\bar{\rho}$ (i.e., $\rho_j \leq \bar{\rho}$) where $0 < \bar{\rho} < 1$ is a constant. Therefore, for a given $0 < \bar{\rho} < 1$, we replace the stability condition (4.3) by

$$\bar{\lambda} \sum_{i \in \mathcal{L}} x_{ij} \frac{F}{r_{ij}} \leq \bar{\rho} \ , \quad \forall j \in \mathcal{B} \cup \{0\}. \tag{4.4}$$

Given the resource allocation $\boldsymbol{RA}$, our objective is to compute the highest possible arrival rate $\overline{\lambda}_{max}(\boldsymbol{RA})$ for which there exists an user association such that the system is stable. To compute $\overline{\lambda}_{max}(\boldsymbol{RA})$, we formulate a joint optimization problem in which the variables are the $\{x_{ij}\}$'s and $\bar{\lambda}$. The problem can be formulated as follows: Given $\bar{\rho}$, $\{r_{ij}\}$, and F, compute $\{x_{ij}\}$ and $\bar{\lambda}$ so as to maximize the arrival rate:

$$\begin{aligned}
\mathbf{P}_s^{(1)}: \quad & \max_{\{x_{ij}\}, \bar{\lambda}} \quad \bar{\lambda} \\
\text{subject to} \quad & \bar{\lambda} \sum_{i \in \mathcal{L}} x_{ij} \frac{F}{r_{ij}} \leq \bar{\rho} \ , \quad \forall j \in \mathcal{B} \cup \{0\} && \text{(4.5a)} \\
& \sum_{j \in \mathcal{B} \cup \{0\}} x_{ij} = 1 \ , \quad \forall i \in \mathcal{L} && \text{(4.5b)} \\
& \bar{\lambda} \geq 0, \ x_{ij} \in \{0, 1\} \ , \quad \forall i \in \mathcal{L}, \ \forall j \in \mathcal{B} \cup \{0\} && \text{(4.5c)}
\end{aligned}$$

Note that maximizing $\bar{\lambda}$ is equivalent to minimizing the maximum of $\left(\sum_{i \in \mathcal{L}} x_{ij} \frac{F}{r_{ij}}\right)$'s

[4]Strict inequalities result in open sets.

for all $j \in \mathcal{B} \cup \{0\}$. Using this property, $\mathbf{P}_s^{(1)}$ can be rewritten as follows:

$$\begin{aligned}
\mathbf{P}_s^{(2)}: \quad & \min_{\{x_{ij}\},T} \quad T \\
\text{subject to} \quad & \sum_{i\in\mathcal{L}} x_{ij}\frac{F}{r_{ij}} \leq T \ , \quad \forall j \in \mathcal{B}\cup\{0\} && \text{(4.6a)} \\
& \sum_{j\in\mathcal{B}\cup\{0\}} x_{ij} = 1 \ , \quad \forall i \in \mathcal{L} && \text{(4.6b)} \\
& T \geq 0,\ x_{ij} \in \{0,1\} \ , \quad \forall i \in \mathcal{L},\ \forall j \in \mathcal{B}\cup\{0\} && \text{(4.6c)}
\end{aligned}$$

$\mathbf{P}_s^{(2)}$ is an integer linear program which can be solved with a commercial solver software. Therefore, given $\bar{\rho}$, F, and a resource allocation (i.e., $\boldsymbol{RA}$), the highest arrival rate $\bar{\lambda}_{max}$ is

$$\overline{\lambda}_{max}(\boldsymbol{RA}) = \frac{\bar{\rho}}{T^\star} \tag{4.7}$$

where $T^\star$ denotes the optimal solution to $\mathbf{P}_s^{(2)}$.

The proposed optimization problem enables us to compute the highest possible arrival rate $\overline{\lambda}_{max}(\boldsymbol{RA})$ for a given resource allocation $\boldsymbol{RA}$. Arrival rate $\bar{\lambda}$ can be as large as $\overline{\lambda}_{max}(\boldsymbol{RA})$ if we use the user association $\{x_{ij}^\star\}$ where $\{x_{ij}^\star\}$ is the solution to $\mathbf{P}_s^{(2)}$; otherwise, the system is not necessarily stable. For any other user association rule $\boldsymbol{UA} = \{x_{ij}^o\}$ (i.e., any other feasible solution $\{x_{ij}^o\}$ to $\mathbf{P}_s^{(2)}$), the highest possible arrival rate (call it $\widetilde{\lambda}_{max}(\boldsymbol{RA}, \{x_{ij}^o\})$) can be easily computed by using (4.4), and is less than or equal to $\overline{\lambda}_{max}(\boldsymbol{RA})$. In Section 4.5, we compute the maximum arrival rate $\overline{\lambda}_{max}(\boldsymbol{RA})$ for different $\boldsymbol{RA}$ schemes, and study the stability region of different combinations of user association and resource allocation schemes in more details.

Next, we focus on the network delay metrics. We fix the resource allocation and its

parameter if any, and try to compute the lowest possible values of our delay metrics for a given $\bar{\lambda}$. We formulate a problem of optimal user association assuming the resource allocation $\boldsymbol{RA}$, $\overline{\lambda}_{max}(\boldsymbol{RA})$, and an arrival rate $\bar{\lambda} \leq \overline{\lambda}_{max}(\boldsymbol{RA})$ are given.

4.2.2 Network Delay

Let us assume that a resource allocation, its parameter if any (call the resource allocation $\boldsymbol{RA}$), and an arrival rate $\bar{\lambda}$ which is less than or equal to $\overline{\lambda}_{max}(\boldsymbol{RA})$ (the value computed for the resource allocation), are given. Given the resource allocation and arrival rate $\bar{\lambda}$, our objective is to find the lowest possible values of our delay metrics, i.e., the lowest possible values of $d_1(\{T_i\}) = \max_{i \in \mathcal{L}} T_i$ and $d_2(\{T_i\}) = \frac{1}{|\mathcal{L}|}\sum_{i \in \mathcal{L}} T_i$ where T_i denotes the average delay at location i (i.e., class i). Given the resource allocation, we can compute $\{r_{ij}\}$. Therefore, for a given user association $\boldsymbol{UA} = \{x^o_{ij}\}$, we can compute the average delay at location i [68], i.e.,

$$T^o_i = \sum_{j \in \mathcal{B} \cup \{0\}} x^o_{ij} \frac{F}{(1-\rho_j) r_{ij}} \tag{4.8}$$

where $\rho_j = \bar{\lambda} \sum_{i \in \mathcal{L}} x^o_{ij} \frac{F}{r_{ij}}$ is the load factor of BS j. Hence, we can easily compute the values of our delay metrics when an user association scheme is given. We might also be interested in finding the optimal association, i.e., the association that would yield the minimum delay metric. Then, given the resource allocation (i.e., either CCD, or OD or PSD with a given K) and arrival rate $\bar{\lambda}$, our objective is to find the user associations that minimize our delay metrics (i.e., $d_1(\{T_i\})$ and $d_2(\{T_i\})$). To find the lowest possible values of our delay metrics and the optimal user association, we formulate an optimal user association problem.

Given the resource allocation and arrival rate $\bar{\lambda}$, the problem is to find the user association that minimizes our delay metrics (either $d_1(\{T_i\}) = \max_{i\in\mathcal{L}} T_i$ or $d_2(\{T_i\}) = \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} T_i$). We formulate an optimal user association problem in which the variables are the $\{x_{ij}\}$. The problem can be formulated as follows: Given $\bar{\rho}$, $\bar{\lambda}$, the resource allocation $\boldsymbol{RA}$, the channel gains for the L fixed locations, and F, compute $\{x_{ij}\}$ so as to minimize our objective function:

$$\mathbf{P}_{delay}: \quad \min_{\{x_{ij}\},\{\rho_j\},\{T_i\}} \quad d(\{T_i\})$$

subject to $(4.5a) - (4.5b)$

$$T_i \geq x_{ij}\frac{F}{(1-\rho_j)r_{ij}}\ , \quad \forall i \in \mathcal{L},\ \forall j \in \mathcal{B}\cup\{0\} \tag{4.9a}$$

$$\rho_j = \bar{\lambda}\sum_{i\in\mathcal{L}} x_{ij}\frac{F}{r_{ij}},\ \forall j \in \mathcal{B}\cup\{0\} \tag{4.9b}$$

$$x_{ij} \in \{0,1\}\ , \quad \forall i \in \mathcal{L},\ \forall j \in \mathcal{B}\cup\{0\} \tag{4.9c}$$

where r_{ij}'s can be computed beforehand.

$\mathbf{P}_{delay}$ is an integer generalized linear-fractional program, and hence it is not possible to solve $\mathbf{P}_{delay}$ efficiently as it is. Therefore, we have developed numerical techniques to solve $\mathbf{P}_{delay}$. We propose an algorithm to solve $\mathbf{P}_{delay}$ with the desired precision $\epsilon > 0$ when the objective is to minimize the maximum average delay per class, and propose a simple transformation to transform $\mathbf{P}_{delay}$ into a convex integer program when the objective is to minimize the average delay in the cell area. The proposed techniques are provided in Section 4.3. In Section 4.5, using the proposed techniques, we compute the lowest possible delay for different resource allocation schemes, and compare the delay performance of several combinations of user associ-

ation and resource allocation schemes.

Next, we try to find the highest possible arrival rate when there is a constraint on the maximum average delay per class. We fix the resource allocation and its parameter if any, and formulate a problem of optimal user association assuming the resource allocation $\boldsymbol{RA}$, $\overline{\lambda}_{max}(\boldsymbol{RA})$, and a constraint on the highest possible delay per class are given.

4.2.3 Delay-constrained Maximum Throughput

Let us assume that a resource allocation, its parameter if any (call the resource allocation $\boldsymbol{RA}$), $\overline{\lambda}_{max}(\boldsymbol{RA})$ (the value computed for the resource allocation), and a constraint T on the maximum average delay per class are given. Given the resource allocation and the constraint, our objective is to find the highest possible arrival rate (i.e., $\bar{\lambda}_T(\boldsymbol{RA})$) for which the maximum average delay per class is less than or equal to a certain value $T > 0$, i.e., $d(\{T_i\}) = \max_{i \in \mathcal{L}} T_i \leq T$. Given the resource allocation, we can compute $\{r_{ij}\}$. Therefore, for a given user association $\boldsymbol{UA} = \{x_{ij}^o\}$, we can easily compute the highest arrival rate (call it $\widetilde{\lambda}_T(\boldsymbol{RA}, \boldsymbol{UA})$) by using (4.2) and (4.8). We might also be interested in finding the optimal association, i.e., the association that would yield the largest arrival rate. Then, given the resource allocation scheme (i.e., either CCD, or OD or PSD with a given K) and the delay constraint, our objective is to compute the highest possible arrival rate $\bar{\lambda}_T(\boldsymbol{RA})$ for which there exists an user association such that the system is stable, and that the maximum average delay per class is less than or equal to T. To find $\bar{\lambda}_T(\boldsymbol{RA})$ and the optimal user association, we formulate a joint optimization problem.

Given the resource allocation and the delay constraint, the problem is to find the

user association that maximizes the allowable arrival rate $\bar{\lambda}$. We formulate a joint optimization problem in which the variables are the $\{x_{ij}\}$ and $\bar{\lambda}$. The problem can be formulated as follows: Given $\bar{\rho}$, T, the resource allocation $\boldsymbol{RA}$, $\overline{\lambda}_{max}(\boldsymbol{RA})$, the channel gains for the L fixed locations, and F, compute $\{x_{ij}\}$ and $\bar{\lambda}$, so as to maximize arrival rate $\bar{\lambda}$:

$$\mathbf{P}_T: \quad \max_{\{x_{ij}\},\bar{\lambda}} \quad \bar{\lambda}$$

$$\text{subject to } (4.5a)-(4.5b),\ (4.9a)-(4.9c)$$

$$T_i \leq T,\ \forall i \in \mathcal{L} \tag{4.10a}$$

$$0 \leq \bar{\lambda} \leq \overline{\lambda}_{max}(\boldsymbol{RA}) \tag{4.10b}$$

where r_{ij}'s can be computed beforehand. $\mathbf{P}_T$ is a mixed integer non-linear program, and hence it is not possible to obtain exact solutions to $\mathbf{P}_T$ efficiently as it is. However, the structure of $\mathbf{P}_T$ is such that it can be transformed into an integer linear program. In Section 4.3, we propose a simple technique to transform $\mathbf{P}_T$ into a linear integer program which can be solved with a commercial solver software.

Remark 4.2.1 *Given an $\boldsymbol{RA}$ and arrival rate $\bar{\lambda}$, let d_0 denote the lowest possible value of $d_1(\{T_i\})$. Given a constraint d_0 on the maximum average delay per class, let $\bar{\lambda}_{d_0}(\boldsymbol{RA})$ denote the highest allowable arrival rate for the resource allocation $\boldsymbol{RA}$, i.e., the optimal solution to $\boldsymbol{P}_T$. We can show that $\bar{\lambda}_{d_0}(\boldsymbol{RA}) = \bar{\lambda}$.*

Next, we focus on our solution techniques for solving $\mathbf{P}_{delay}$ and $\mathbf{P}_T$.

4.3 Solution Techniques

As mentioned earlier, the proposed problems $\mathbf{P}_{delay}$ and $\mathbf{P}_T$ are mixed integer non-linear programs which are hard to solve in their current forms. We propose simple solution techniques some of which enable us to solve these problems and some others allow us to obtain lower bounds on the system's performance.

4.3.1 Minimizing The Maximum Average Delay per Class

Let us focus on the problem $\mathbf{P}_{delay}$ assuming the objective is to minimize the maximum average delay per class, i.e., $d_1(\{T_i\}) = \max_{i\in\mathcal{L}} T_i$. This problem is an integer generalized linear-fractional program. Unlike linear fractional programs, generalized linear fractional programs cannot be reduced to linear programs. However, we can check whether the optimal value of $\mathbf{P}_{delay}$ is less than or more than a given value t by solving the following feasibility problem: Given $t > 0$, $\bar{\rho}$, $\bar{\lambda}$, $\{r_{ij}\}$, and F, compute $\{x_{ij}\}$ so as to minimize our objective function:

$$\begin{aligned} \mathbf{P}^{(f)}_{delay}(t): \quad & \min_{\{x_{ij}\},\{\rho_j\},\{T_i\}} \quad \mathbf{1} \\ & \text{subject to } (4.5a)-(4.5b),\ (4.9a)-(4.9c) \\ & T_i \le t, \quad i \in \mathcal{L} \end{aligned}$$

Note that $\mathbf{P}^{(f)}_{delay}(t)$ can be rewritten as follows:

$$\begin{aligned} \mathbf{P}^{(f)}_{delay}(t): \quad & \min_{\{x_{ij}\},\{\rho_j\},\{T_i\}} \quad \mathbf{1} \\ & \text{subject to } (4.5a)-(4.5b),\ (4.9b)-(4.9c) \end{aligned}$$

$$t(1-\rho_j) \geq x_{ij}\frac{F}{r_{ij}} \quad \forall i \in \mathcal{L},\ \forall j \in \mathcal{B} \cup \{0\} \tag{4.12a}$$

The feasibility problem $\mathbf{P}^{(f)}_{delay}(t)$ is an integer linear program which can be solved with a commercial solver software.

If the feasibility problem $\mathbf{P}^{(f)}_{delay}(t)$ is feasible, then we have $p^\star \leq t$ where $p^\star$ denotes the optimal value of the objective function in $\mathbf{P}_{delay}$; otherwise, we have $p^\star > t$. We use this observation as the basis of a simple algorithm for solving $\mathbf{P}_{delay}$ using *bisection*. In this algorithm, we first try to find a feasible solution to $\mathbf{P}_{delay}$ using one of the simple association rules which will be introduced in Section 4.4. Let t_0 denote the value of $\max_{i\in\mathcal{L}} T_i$ for the association rule. Let us assume that $\bar{\lambda} \leq \overline{\lambda}_{max}(\boldsymbol{RA})$ (i.e., $\mathbf{P}_{delay}$ is feasible), and start with the interval $I_0 = [0, t_0]$. Clearly, the interval I_0 contains the optimal value of the objective function in $\mathbf{P}_{delay}$. We solve the feasibility problem $\mathbf{P}^{(f)}_{delay}(t)$ at the midpoint of I_0, i.e., $t = \frac{t_0}{2}$. This determines whether the optimal value $p^\star$ is in the lower or upper half of I_0. Then, we obtain a new interval which contains the optimal value $p^\star$ (i.e., the optimal value of the objective function). Note that the width of the new interval is reduced to half of the interval in the previous iteration. We repeat this process until the width of the interval is sufficiently small. In each step, the width of the interval is reduced by two folds, and hence after k iterations, the length of the interval is $2^{-k}t_0$. Therefore, we need $\lceil \log_2(\frac{t_0}{\epsilon}) \rceil$ iterations to obtain the optimal value of $\mathbf{P}_{delay}$ with the desired precision $\epsilon > 0$. A formal description of the bisection algorithm is given in Fig. 4.1.

Unfortunately, the bisection technique is not useful for solving $\mathbf{P}_{delay}$ when the objective is to minimize the average delay in the cell area. In this case, the feasibility problem is a non-linear integer program which is hard to solve. Next, we develop a

Given $t_0 > 0$, tolerance $\epsilon > 0$
Set $\ell = 0$, $u = t_0$, $i = 0$
repeat

1. $t_i := \frac{\ell+u}{2}$
2. Solve the feasibility problem $\mathbf{P}^{(f)}_{delay}(t_i)$.
3. If $\mathbf{P}^{(f)}_{delay}(t_i)$ is feasible, set $u := t_i$; otherwise, set $\ell := t_i$.
4. $i := i + 1$

until $u - \ell \leq \epsilon$

Figure 4.1: A formal description of the bisection algorithm.

simple technique to obtain lower bounds on the lowest average delay in the cell area.

4.3.2 Minimizing The Average Delay in The Cell

Let us focus on $\mathbf{P}_{delay}$, and assume that the objective is to minimize the average delay in the cell area, i.e., $d_2(\{T_i\}) = \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} T_i$. The problem $\mathbf{P}_{delay}$ is a non-convex quadratic program which is hard to solve as it is. Our goal is to obtain lower bounds for $\mathbf{P}_{delay}$. To compute a lower bound for $\mathbf{P}_{delay}$, we could simply relax the integrality constraints on $\{x_{ij}\}$, and try to solve the relaxed problem. However, even after relaxing the integrality constraints in $\mathbf{P}_{delay}$, the problem remains non-convex. Fortunately, the problem $\mathbf{P}_{delay}$ can be reformulated into an integer convex problem.

Using (4.8), we can easily show that the average delay in the cell is equal to $\frac{1}{\bar{\lambda}|\mathcal{L}|}\sum_{j\in\mathcal{B}\cup\{0\}} \frac{\rho_j}{1-\rho_j}$ where $\rho_j = \bar{\lambda}\sum_{i\in\mathcal{L}} x_{ij}\frac{F}{r_{ij}}$ for all j. Therefore, minimizing the average delay in the cell is equivalent to minimizing $\sum_{j\in\mathcal{B}\cup\{0\}} \frac{\rho_j}{1-\rho_j}$ since $\bar{\lambda}$ and $|\mathcal{L}|$ are given beforehand. We can show that $\sum_{j\in\mathcal{B}\cup\{0\}} \frac{\rho_j}{1-\rho_j} = -(|\mathcal{B}|+1) + \sum_{j\in\mathcal{B}\cup\{0\}} \frac{1}{1-\rho_j}$. Hence, $\mathbf{P}_{delay}$ can be rewritten as follows: Given $\bar{\rho}$, $\bar{\lambda}$, $\{r_{ij}\}$, and F, compute $\{x_{ij}\}$ so

as to minimize our objective function:

$$\mathbf{P}'_{delay}: \quad \min_{\{x_{ij}\},\{\rho_j\}} \quad \sum_{j\in\mathcal{B}\cup\{0\}} \frac{1}{1-\rho_j}$$

subject to $(4.5a) - (4.5b)$

$$\rho_j = \bar{\lambda} \sum_{i\in\mathcal{L}} x_{ij} \frac{F}{r_{ij}}, \ \forall j \in \mathcal{B}\cup\{0\} \tag{4.13a}$$

$$x_{ij} \in \{0,1\} \ , \quad \forall i \in \mathcal{L}, \ \forall j \in \mathcal{B}\cup\{0\} \tag{4.13b}$$

where r_{ij}'s can be computed beforehand. The problem $\mathbf{P}'_{delay}$ is a convex integer program, and its relaxed program can be solved efficiently even for large systems. Note that convex programs can be solved to the desired precision in polynomial time [54]. This enables us to obtain lower bounds on the lowest average delay in the cell area.

We will numerically verify the tightness of the computed lower bound (i.e., the solution to $\mathbf{P}'_{delay}$) by finding a feasible solution to the problem $\mathbf{P}_{delay}$, and then comparing the average delay $d_2(\{T_i\}) = \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} T_i$ for this feasible solution with the solution to $\mathbf{P}'_{delay}$. Note that we will use simple association rules to generate feasible solutions for a given resource allocation. In this case, the average delay can be determined easily.

4.3.3 Maximizing The Delay-Constrained Throughput

The problem $\mathbf{P}_T$ is a non-convex quadratic program which is hard to solve as it is. This is due to the quadratic constraint (4.5a). Our goal is to transform the problem into a linear integer program at the cost of additional variables and constraints. To

do so, we will need two steps. In the first step, we make sure that there exists a feasible solution to $\mathbf{P}_T$ with a non-zero arrival rate. Clearly, if there does not exist such a feasible solution, then the delay-constrained maximum throughput is zero (i.e., $\bar{\lambda}_T(\boldsymbol{RA}) = 0$). Otherwise, we proceed to the second step in which we transform $\mathbf{P}_T$ into a linear integer program (we call it $\mathbf{P}'_T$), and compute the delay-constrained maximum throughput $\bar{\lambda}_T(\boldsymbol{RA})$.

STEP 1 : Given $\bar{\rho}$, T, $\overline{\lambda}_{max}(\boldsymbol{RA})$, $\{r_{ij}\}$, and F, we can verify that $\mathbf{P}_T$ has a feasible solution with a non-zero arrival rate *if and only if* for each location $i \in \mathcal{L}$, there exists a BS $j^\star$ such that $F < Tr_{ij^\star}$. Using this property, we find that the delay-constrained maximum throughput $\bar{\lambda}_T(\boldsymbol{RA})$ is zero if there exists a location i for which $F \geq Tr_{ij}$ for all $j \in \mathcal{B} \cup \{0\}$; otherwise $\bar{\lambda}_T(\boldsymbol{RA}) > 0$.

STEP 2 : Let us assume that the delay-constrained maximum throughput is not zero (i.e., $\bar{\lambda}_T(\boldsymbol{RA}) > 0$). We define a new variable $\lambda' = 1/\bar{\lambda}$, and reformulate $\mathbf{P}_T$ by using the new variable. $\mathbf{P}_T$ can be rewritten as follows:

$$\mathbf{P}_T : \quad \min_{\{x_{ij}\},\lambda'} \quad \lambda'$$

subject to

$$\rho_j = \frac{1}{\lambda'} \sum_{i \in \mathcal{L}} x_{ij} \frac{F}{r_{ij}}, \ \forall j \in \mathcal{B} \cup \{0\} \tag{4.14a}$$

$$\lambda' x_{ij} \frac{F}{Tr_{ij}} \leq \left(\lambda' - \sum_{i \in \mathcal{L}} x_{ij} \frac{F}{r_{ij}} \right), \quad \forall i \in \mathcal{L}, \ \forall j \in \mathcal{B} \cup \{0\} \tag{4.14b}$$

$$\sum_{j \in \mathcal{B} \cup \{0\}} x_{ij} = 1 \ , \quad \forall i \in \mathcal{L} \tag{4.14c}$$

$$\frac{1}{\overline{\lambda}_{max}(\boldsymbol{RA})} \leq \lambda' \tag{4.14d}$$

$$x_{ij} \in \{0,1\} \ , \quad \rho_j \leq \bar{\rho} \quad \forall i \in \mathcal{L},\ \forall j \in \mathcal{B} \cup \{0\} \tag{4.14e}$$

Although the new problem is not a linear integer problem, the mixed quadratic terms involving binary variables (i.e., $\lambda' x_{ij}$s) can be linearized (exactly) with the linearization techniques proposed in [69].

Proposition 4.3.1 *[69] Given a binary variable x and any (linear) function $g(w)$ in a continuous variable $w \in W$ (bounded), a variable z equals the quadratic function $xg(w)$ if and only if*

$$Lx \leq z \leq Ux, \text{ and } g(w) - U(1-x) \leq z \leq g(w) - L(1-x)$$

where $L = \min\{g(w) : w \in W\}$ and $U = \max\{g(w) : w \in W\}$ assumed finite.

Let's define a new variable λ'_{ij} for each pair (i,j) where $i \in \mathcal{L}$ and $j \in \mathcal{B} \cup \{0\}$:

$$\lambda'_{ij} = \lambda' \ x_{ij} \ .$$

Note that a lower bound on λ' is given in (4.14d). Therefore, we only need to find an upper bound on λ', i.e., a lower bound on $\bar{\lambda}$. Since $\bar{\lambda}_T(\boldsymbol{RA}) \neq 0$, there exists a strictly positive lower bound (we call it $\bar{\lambda}_0$) on $\bar{\lambda}_T(\boldsymbol{RA})$. To find such a lower bound, we associate each location $i \in \mathcal{L}$ with a BS $j^\star$ for which $F < Tr_{ij^\star}$, and use the problem $\mathbf{P}_T$ to compute $\bar{\lambda}_0$. Note that the problem $\mathbf{P}_T$ can be used to compute the delay-constrained maximum throughput for a given association rule. It can be verified that the computed value $\bar{\lambda}_0$ is strictly positive, i.e., $(1/\bar{\lambda}_0)$ is an upper bound on λ'.

Using Proposition 4.3.1, $\mathbf{P}_T$ can be reformulated as follows:

$$\mathbf{P}'_T: \quad \min_{\{x_{ij}\},\lambda'} \quad \lambda'$$

subject to

$$\sum_{i\in\mathcal{L}} x_{ij}\frac{F}{r_{ij}} \leq \lambda'\bar{\rho},\ \forall j \in \mathcal{B}\cup\{0\} \tag{4.15a}$$

$$\lambda'_{ij}\frac{F}{Tr_{ij}} \leq \left(\lambda' - \sum_{i\in\mathcal{L}} x_{ij}\frac{F}{r_{ij}}\right),\quad \forall i \in \mathcal{L},\ \forall j \in \mathcal{B}\cup\{0\} \tag{4.15b}$$

$$\sum_{j\in\mathcal{B}\cup\{0\}} x_{ij} = 1\ ,\quad \forall i \in \mathcal{L} \tag{4.15c}$$

$$\frac{x_{ij}}{\overline{\lambda}_{max}(\boldsymbol{RA})} \leq \lambda'_{ij} \leq \frac{x_{ij}}{\bar{\lambda}_0} \tag{4.15d}$$

$$\lambda' - \frac{(1-x_{ij})}{\bar{\lambda}_0} \leq \lambda'_{ij} \leq \lambda' - \frac{(1-x_{ij})}{\overline{\lambda}_{max}(\boldsymbol{RA})} \tag{4.15e}$$

$$x_{ij} \in \{0,1\}\ ,\quad \forall i \in \mathcal{L},\ \forall j \in \mathcal{B}\cup\{0\} \tag{4.15f}$$

$\mathbf{P}'_T$ is a linear integer program which can be solved with a commercial solver software.

The proposed user association problems enable us to compute the optimal values of our performance metrics (i.e., the highest possible arrival rate, the maximum average delay per class, the average delay in the cell, and the delay-constrained maximum throughput) for a given resource allocation (i.e., either *CCD* or, *OD* or *PSD* with a given K). We can find the optimal resource allocation and its parameter if any, by computing the optimal values of our performance metrics iteratively for all possible values of K, and then selecting the best solution. Using this technique, we can compare different combinations of user association and resource allocation schemes. Next, we describe the simple association rules that we are going to study.

4.4 User Association Rules

In our system, we assume that users arrive, and depart when their files have been downloaded completely. To operate this system optimally, we need to compute the optimal resource allocation and user association whenever the user arrival rate $\bar{\lambda}$ or channel gains change. It is hard to do such heavy computations online. In addition, changing the resource allocation and its parameters if any, and re-associating users too often might degrade the system's performance. To avoid such problems, network operators associate users in a distributed fashion using some simple user association rules. Typically, these rules use physical layer parameters such as received signal power, path loss, and SINR, to determine the BS each user should associate to. We study the simple user association rules introduced in Chapter 3:

1. **Best SINR:** A user at location $i \in \mathcal{L}$ associates with BS $j^\star$ that provides the highest SINR, i.e., $j^\star = \arg\max_{j \in \mathcal{B} \cup \{0\}} \{\gamma_{ij}^{(c)}\}$ where $\gamma_{ij}^{(c)}$ denotes the SINR of location i from BS j, on each sub-channel respectively. This association rule has been used in homogeneous cellular networks.

2. **Range Extension (RE) [55]:** A user at location i associates with BS $j^\star = \arg\min_{j \in \mathcal{B} \cup \{0\}} \{\delta_{ij}\}$ where δ_{ij} is the path loss from BS j to location i.

3. **Picocell First (PCF) [50]:** A user at location i associates with pico BS $j^\star = \arg\max_{j \in \mathcal{B} \cup \{0\}} \{\gamma_{ij}^{(c)}\}$ as long as $\gamma_{ij^\star}^{(c)} > \beta$ where β is a tuning parameter. Note that $\gamma_{ij}^{(c)}$ denotes the SINR at location i on each sub-channel. If $\max_{j \in \mathcal{B}} \{\gamma_{ij}^{(c)}\} < \beta$, the user at location i associates with the BS that gives the maximum SINR.

For each of these rules, we can compute the values of x_{ij} for all locations i and BSs j when we fix the resource allocation scheme and its parameters if any (i.e., either *CCD* or, *OD* or *PSD* with a given K). Hence, given a combination of a resource allocation and user association rule, we can easily compute the system's performance (i.e., the highest possible arrival rate, the maximum average delay per class, the average delay in the cell, and the delay-constrained maximum throughput). We can also find the optimal resource allocation and its parameter if any, by computing the optimal values of our performance metrics iteratively for all possible values of K, and then selecting the best solution. Using this technique, we can compare different combinations of user association and resource allocation schemes. Next, we explore the performance of existing user association and resource allocation schemes.

4.5 Numerical Results

4.5.1 Parameter Settings

We consider a system composed of 19 macro BSs. Each macro BS's coverage area is overlaid with four pico BSs. The system has an inter-cell distance of 500 m. We study the cell at the center which is a hexagonal area with radius $R = 500/\sqrt{3}\ m$ (see Fig. 4.2). The macro BS is located at the center of the cell while the pico BSs are located around the macro BS with distance $d = 230\ m$ symmetrically from the center. As mentioned earlier, we assume that the system is an OFDM system with M' sub-channels. We consider a reuse factor of "three", i.e., each macro BS has access to $M = M'/3$ sub-channels. We use $M = 100$ sub-channels, and take $\bar{\rho} = 0.95$.

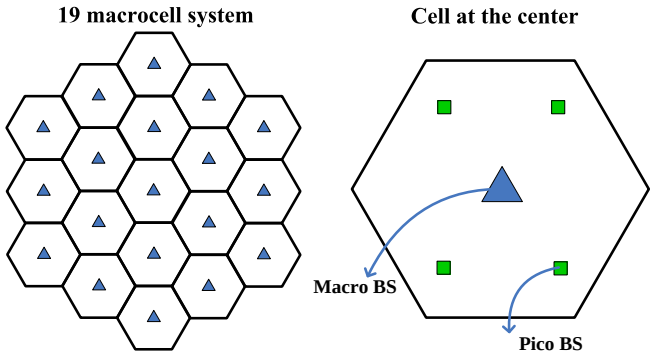


Figure 4.2: A Hetnet comprising 19 macro BSs (the triangles), and many pico BSs (the squares). Each macro BS is overlaid with 4 pico BSs. Pico locations for the cell at the center are shown in the right-hand side figure. The triangle is the macro BS and the squares are the pico BSs.

Table 4.1: Physical Layer Parameters

Noise Power	$-174\ \frac{\text{dBm}}{\text{Hz}}$	T_{subframe}	1 ms
P_{pico}	30 dBm	P_{macro}	46 dBm
UE Ant. Gain	0 dB	Sub-channel Bandwidth	180 KHz
Shadowing s.d.	8 dB	Penetration Loss	20 dB
SC_{ofdm}	12	SY_{ofdm}	14
Path Loss Pico	$140.7 + 36.7\ \log_{10}(d/1000),\ d \geq 10m$		
Path Loss Macro	$128 + 37.6\ \log_{10}(d/1000),\ d \geq 35m$		

We assume that there are $L = 2000$ possible user locations in the cell area. Users arrive into each location according to a Poisson process with density $\bar{\lambda}$, i.e., users arrive into the cell area according to a Poisson process with density $\lambda = L\bar{\lambda}$. We assume that users arriving to the system download files whose sizes are independent and identically distributed (i.i.d.) random variables of mean $F = 10^6$ bits.

The physical layer parameters are based on the 3GPP evaluation methodology document [56] used for Hetnets in LTE. These parameters are shown in Table 4.1. We use the SINR model introduced in Section 4.1 that accounts for path loss and slow

Table 4.2: Modulation and Coding Schemes-LTE

SINR thresholds (in dB)	-6.5	-4	-2.6	-1	1	3	6.6	10	11.4	11.8	13	13.8	15.6	16.8	17.6
Efficiency (in bits/symbol)	0.15	0.23	0.38	0.60	0.88	1.18	1.48	1.91	2.41	2.73	3.32	3.90	4.52	5.12	5.55

Table 4.3: The SINR threshold values for the tuning parameter β

β	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
SINR thresholds (in dB)	-6.5	-4	-2.6	-1	1	3	6.6	8.17	9.33	10.24

fading. Slow fading is modeled by a log-normal shadowing with standard deviation of 8 dB, and path losses for pico and macro BSs are given in Table 4.1. We assume that the system uses adaptive modulation and coding with discrete rates. Table 4.2 taken from [57] and [58] gives us the mapping between the SINR and the efficiency (in bits/symbol) for the modulation and coding schemes (MCS) for LTE. The bit rate obtained by a user that has a SINR between level ℓ and level $\ell+1$ is $r = \frac{\mathrm{SC}_{\mathrm{ofdm}}\ \mathrm{SY}_{\mathrm{ofdm}}}{T_{\mathrm{subframe}}} e_\ell$ where e_ℓ is the efficiency (bits/symbol) of the corresponding level ℓ, $\mathrm{SC}_{\mathrm{ofdm}}$ is the number of data subcarriers per sub-channel bandwidth, $\mathrm{SY}_{\mathrm{ofdm}}$ is the number of OFDM symbols per subframe, and T_{subframe} is the subframe duration in time units. The value of these parameters are shown in Table 4.1. "Picocell First" has a tuning parameter β . We assume that β can take any one of the SINR threshold values shown in Tables 4.3-4.4.

As mentioned earlier in Remark 4.2.1, the delay performance metrics and the delay-constrained maximum throughput are closely related to each other. In this section, we focus on the highest possible arrival rate, the maximum average delay

Table 4.4: The SINR threshold values for the tuning parameter β

β	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
SINR thresholds (in dB)	10.99	11.63	12.19	12.68	13.13	13.53	13.8	15.6	16.8	17.6

per class, and the average delay in the cell, and provide numerical results on these metrics. We compute the optimal values of our performance metrics (i.e., the highest possible arrival rate, the maximum average delay per class, and the average delay in the cell) for different resource allocation schemes as well as the values of our metrics for different combinations of UA and RA schemes for 20 networks. A network corresponds to the random realization of the shadowing coefficients for the $L = 2000$ locations from all the BSs in the multi-tier system. In contrast to the static modeling approach, we do not need to randomly drop users in the system area, and compute the average results over multiple realizations. We only need to consider multiple network realizations corresponding to different shadowing environments. In this section, we show the trends averaged over the 20 networks.

4.5.2 Comparison Results

We now focus on *Network Stability*, and compare different resource allocation schemes in terms of the highest possible arrival rate. We fix the resource allocation and its parameters if any, i.e., K for OD or PSD, and compute the highest possible arrival rate $\overline{\lambda}_{max}(\boldsymbol{RA})$ for different RA schemes (i.e., *PSD*, *OD*, and *CCD*) as well as $\widetilde{\lambda}_{max}(\boldsymbol{RA}, \boldsymbol{UA})$ for different combinations of UA and RA schemes. More precisely, we fix the resource allocation (we call it **RA**), and compute the solution to the problem

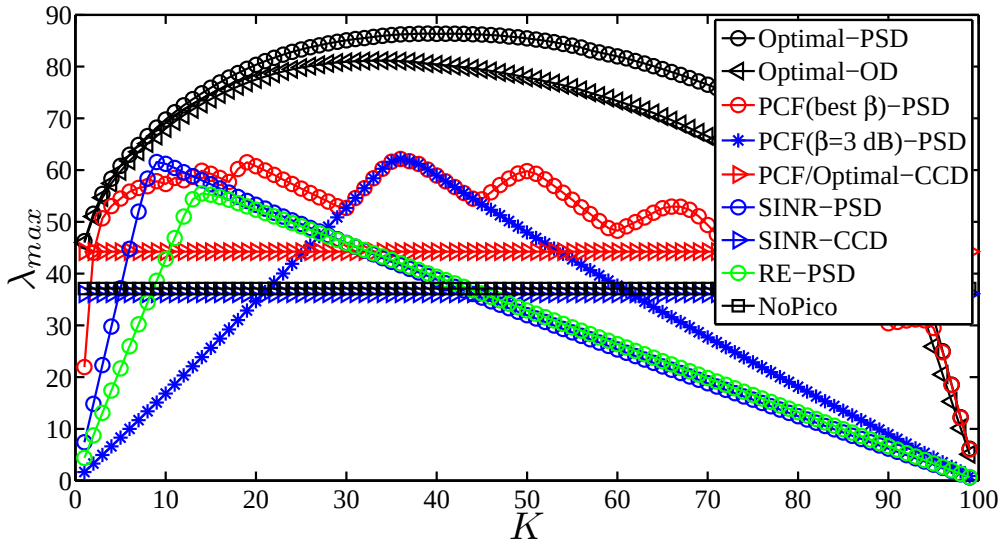


Figure 4.3: PSD, OD, and CCD: The highest user arrival rate as a function of K when $\bar{\rho} = 0.95$, and $F = 10^6$ bits. For each K, we choose the best β. Note that the results are averaged over 20 networks.

$\mathbf{P}_s^{(2)}$ and the corresponding highest arrival rate $\widetilde{\lambda}_{max}(\boldsymbol{RA}, \boldsymbol{UA})$ for different UA rules. Fig. 4.3 shows the highest possible arrival rate $\overline{\lambda}_{max}(\boldsymbol{RA})$ for different RA schemes (i.e., *PSD*, *OD*, and *CCD*) as well as $\widetilde{\lambda}_{max}(\boldsymbol{RA}, \boldsymbol{UA})$ for different combinations of UA and RA schemes. We also show the highest possible arrival rate for the system without pico BSs. The curves corresponding to the optimal solution to $\mathbf{P}_s^{(2)}$ (corresponding to the optimal user association) are labeled *Optimal* in the figure. The results show that:

- *PSD* and *OD* work significantly better than *CCD* for almost all values of K. *PSD* and *OD* perform almost the same with a slight advantage for *PSD*. For *PSD*, we saw an average gain of 100% (with respect to *CCD*) for the highest arrival rate. *CCD* cannot improve the highest arrival rate significantly (with

respect to the system without pico BSs). It can even degrade the highest arrival rate if the operator does not choose the right UA scheme.

- For *CCD*, the association rules *Picocell First* and *Range Extension* are almost optimal since the highest possible arrival rates of the rules *Picocell First* and *Range Extension* are very close to the optimal arrival rate for *CCD*. However, the association rule *Best SINR* does not perform very well since its highest arrival rate is strictly less than the highest arrival rate in the system without any pico BSs.

- For *PSD* and *OD*, the association rules *Picocell First*, *Range Extension*, and *Best SINR* perform almost the same with a slight advantage for *Picocell First*. The performance of these rules is far from the optimal. This can be explained by the fact that load balancing plays a critical role and none of our rules take load balancing into account.

- For *PSD* and *OD*, the association rule *Picocell First* performs almost the same for a large range of values of K when the tuning parameter β is optimized. However, all the association rules *Range Extension*, *Best SINR*, and *Picocell First* with a fixed β are sensitive to the value of the resource allocation parameter K. If the value of the parameter K is not chosen carefully, all the simple rules (i.e., the non-optimal rules) can do worse than the system without pico BSs.

- The comparison of the highest arrival rate (using the optimal solution) between the system with and without pico BSs ("No Pico" in the figure) shows that pico BSs can significantly increase the highest arrival rate. We saw gains (with

respect to the system without pico BSs) in the highest possible arrival rate in the range of 110% to 130%.

4.5.3 In Depth Study of PSD

We now study *partially shared deployment* in more details[5]. To do so, we first focus on the maximum average delay per class, and then study the average delay in the cell area.

Maximum average delay per class:

We select the maximum average delay per class as our delay metric, and try to minimize it, i.e., we try to minimize $\max_{i\in\mathcal{L}} T_i$ where T_i denotes the average delay for class i. We compare the delay performance of the simple UA rules with the optimal delay performance as a function of λ. We fix the resource allocation parameter K as well as the arrival rate λ, and compute the optimal solution to the problem $\mathbf{P}_{delay}$ with the desired precision $\epsilon = 0.02$, and the corresponding maximum average delay per class for each UA rule. For each value of λ, we obtain the optimal delay performance by solving $\mathbf{P}_{delay}$ iteratively for all values of K, and then selecting the best solution. Similarly, for each UA rule, we select the value of K which results in the lowest maximum average delay per class. The results for two non-overlapping ranges of λ are shown in Figures 4.4-4.5. The curve corresponding to the optimal solution is labeled *Optimal* in the figures. The results show that:

- The association rules *Picocell First* and *Best SINR* perform almost the same

[5]We have obtained similar results for *OD*.

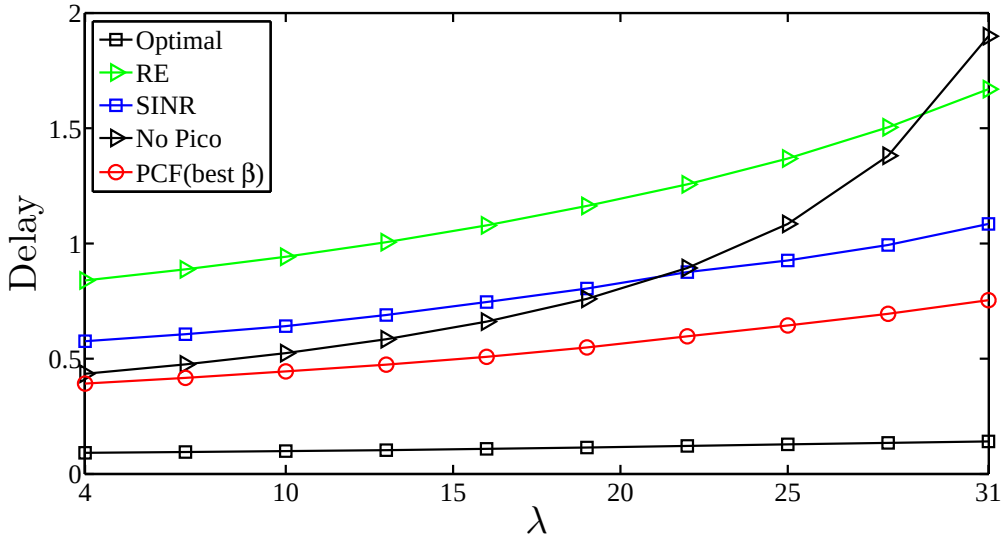


Figure 4.4: PSD: The maximum average delay per class as a function of λ when $F = 10^6$ bits. We choose the best values of K and β. Note that the results are averaged over 20 networks.

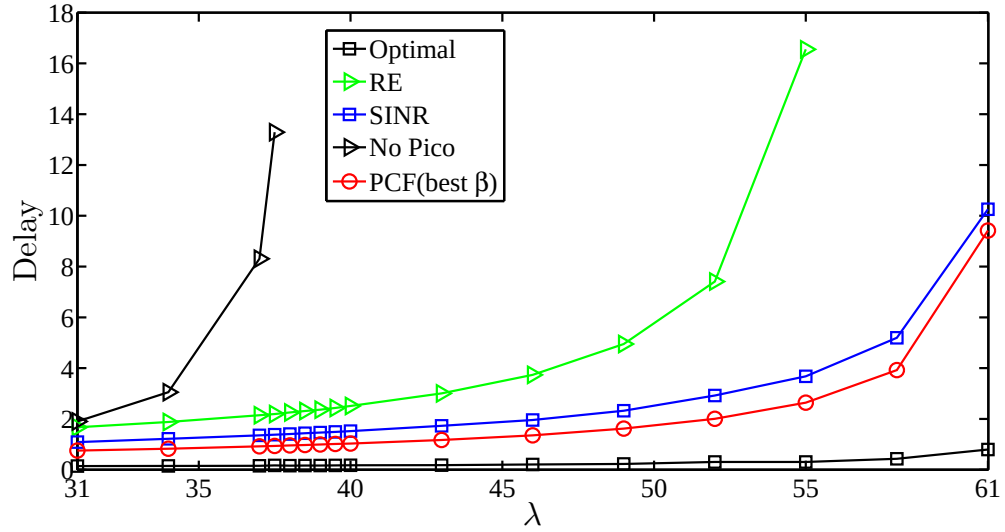


Figure 4.5: PSD: The maximum average delay per class as a function of λ when $F = 10^6$ bits. We choose the best values of K and β. Note that the results are averaged over 20 networks.

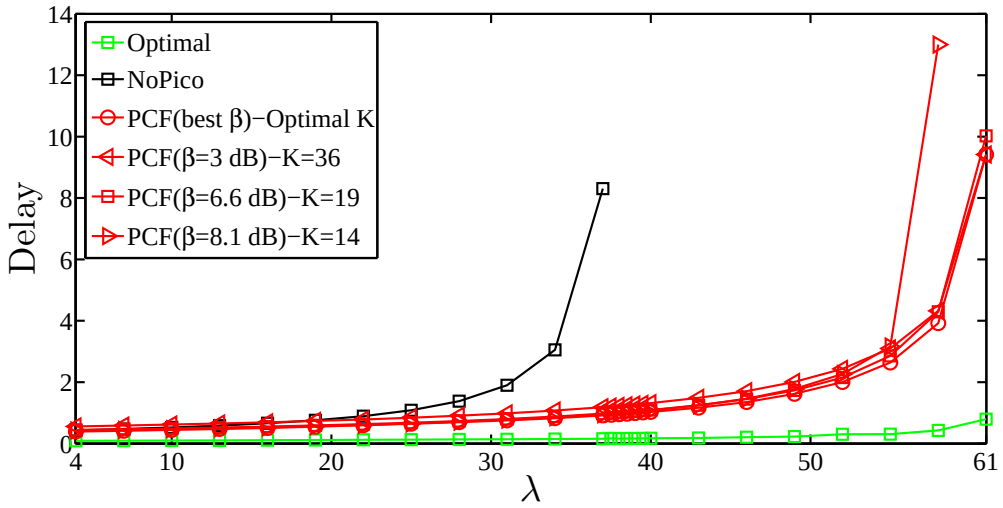


Figure 4.6: PSD, PCF(β): The maximum average delay per class as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

with a slight advantage for *Picocell First*, and they work significantly better than *Range Extension* for all values of λ when we select β and K carefully.

- The association rule *Picocell First* performs better than the system without pico BSs for all values of λ. However, the rules *Best SINR* and *Range Extension* do not always perform better than the system without pico BSs (especially for low values of λ).

- None of the simple rules are performing very well for high values of λ, i.e., they are quite far from the optimal user association.

- The association rule *Range Extension* is not performing better than the system without pico BSs for low values of λ while it is performing better than the system without pico BS for high values of λ.

We now study the impact of the RA parameter K as well as the tuning parameter β on the performance of *Picocell First*. To do this, we compare the delay performance of *Picocell First* as a function of λ for different values of K and β. We select three values of β for which *Picocell First* performs relatively well for all values of λ when we optimize K. The values of β are β_6, β_7, and β_8 taken from Table 4.3. Each value of β corresponds to an instantiation of the *PCF* rule. For each value of β, we select the value of K that is optimal for $\lambda = 61$ users per second (the highest arrival rate that we consider). Given the values of β and K, we fix the arrival rate λ, and compute the maximum average delay per class. We also compute the optimal delay performance as well as the delay performance of *Picocell First* when the parameters K and β are optimized. The results provided in Fig. 4.6 show that:

- The association rules corresponding to β_6, β_7, and β_8 perform almost the same with a slight advantage for $\beta_6 = 3dB$. The performance of *Picocell First* is not very sensitive to the value of β.

- Given a value of β, the value of K that is optimal for λ_1, is quasi-optimal for all values of λ which are smaller than λ_1.

Average delay in the cell:

We compare the average delay of the simple UA rules with the lower bound on the optimal average delay as a function of λ. We fix the arrival rate λ, and compute the lower bound of the optimal user association problem $\mathbf{P}_{delay}$ (i.e., the relaxed solution to $\mathbf{P}_{delay}$) for *PSD*, and the average delay for each UA rule when the RA parameter K is computed optimally. To check the tightness of the computed lower bound, we

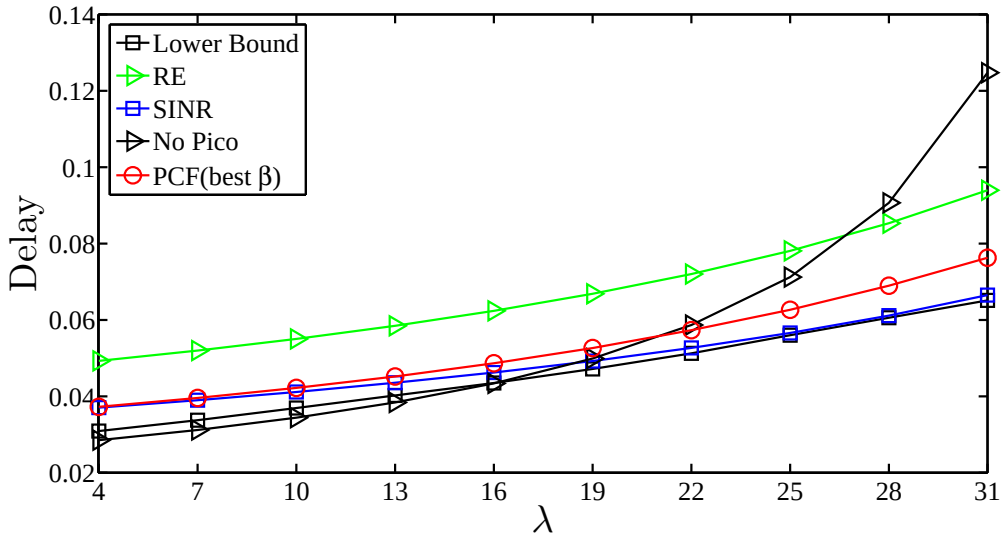


Figure 4.7: PSD: The average delay in the cell as a function of λ when $F = 10^6$ bits. We choose the best values of K and β. Note that the results are averaged over 20 networks.

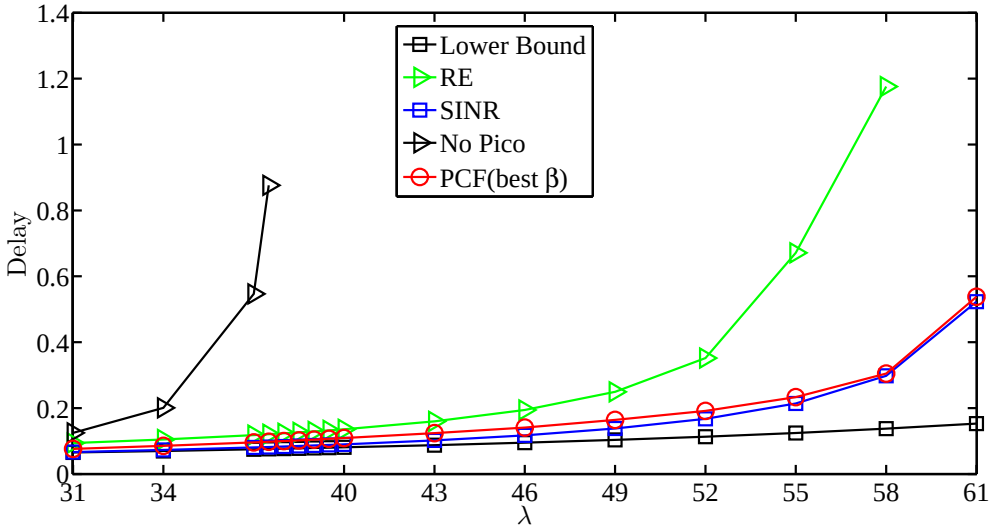


Figure 4.8: PSD: The average delay in the cell as a function of λ when $F = 10^6$ bits. We choose the best values of K and β. Note that the results are averaged over 20 networks.

compare it with the average delay of the simple UA rules. The results for two non-overlapping ranges of λ are shown in Figures 4.7-4.8. The curve corresponding to the lower bound is labeled *Lower Bound* in the figures. The results show that:

- The UA rule *Best SINR* is almost optimal since the average delay of *Best SINR* is very close to the computed lower bound for a large range of λ when the RA parameter K is chosen optimally. This validates our relaxation approach since an integer solution to the proposed problem obtains almost the same average delay as the solution of the relaxed problem. This observation also shows that *Best SINR* is a good UA rule for minimizing the average delay in the cell area if we choose K optimally.

- The association rules *Picocell First* and *Best SINR* perform almost the same with a sight advantage for *Best SINR*, and they work significantly better than *Range Extension* for all values of λ when we select β and K carefully.

- The comparison of the delay performance (using *the lower bound*) between the system with and without pico BSs shows that pico BSs can increase the average delay in the cell when the arrival rate is relatively low (less than 19 users per second). This shows the critical impact of the interference caused by the pico BSs on the delay performance for small values of λ.

We now try to understand the impact of the system parameters K and β on the average delay of *Picocell First*. We take β_6 and β_7 from Table 4.3, and we compare the delay performance of *Picocell First* as a function of λ for different values of K and β. Each value of β corresponds to an instantiation of the *PCF* rule. For each value of β, we select the value of K that is optimal for $\lambda = 61$ users per second. Our

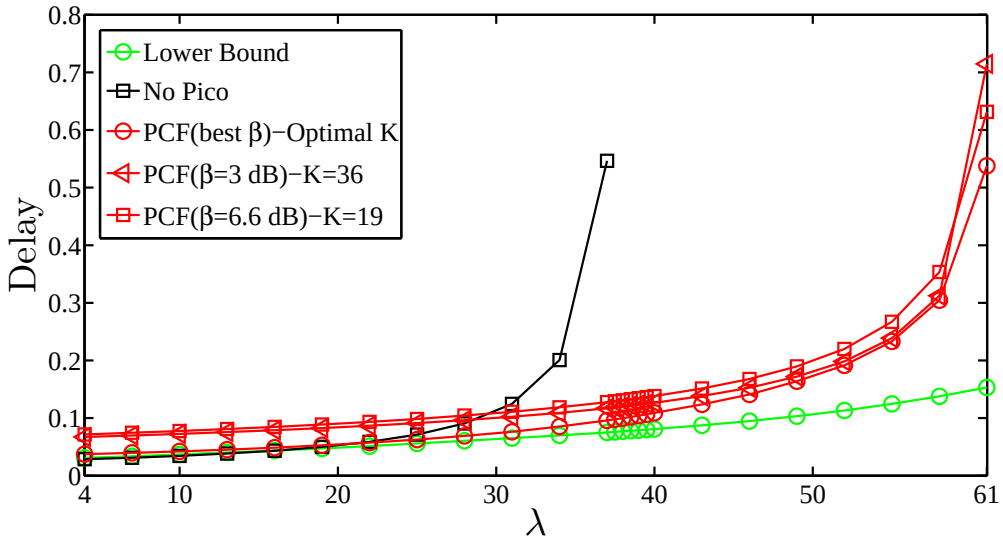


Figure 4.9: PSD, Best SINR, PCF(β): The average delay in the cell as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

numerical results are shown in Fig. 4.9. We also show the lower bound on the average delay as well as the average delay of *Picocell First* when the parameters K and β are chosen optimally. The results show that:

- The association rules corresponding to β_6 and β_7 perform almost the same with a slight advantage for β_6.
- Given a value of β, the value of K that is optimal for λ_1, is quasi-optimal for all values of λ which are smaller than λ_1.

4.6 Conclusions

We have studied the problem of joint user association and resource allocation in Hetnets in which users arrive in the system, stay while downloading the file they

need, and then depart. We have considered three resource allocation schemes *CCD*, *OD*, and *PSD*, and three different performance metrics (the highest possible arrival rate, the network delay, and the delay-constrained maximum throughput). Given a resource allocation and its parameter, we have formulated three optimal user association problems to optimize our performance metrics. The proposed problems are non-linear integer programs, and hence it is not possible to efficiently obtain optimal solutions. We have developed numerical techniques to compute either the exact solutions or tight lower bounds to these problems. Our numerical results show that the proposed lower bounds are tight and can be used as benchmarks to analyze and compare different combinations of user association and resource allocation schemes.

Our numerical results show that pico BSs can improve the system's performance if the network operator uses the right combination of user association and resource allocation schemes; otherwise, pico BSs can degrade the system's performance, and increase the average delay in the cell significantly. The results indicate that *partially shared deployment* and *orthogonal deployment* perform significantly better than *co-channel deployment* if we associate users optimally. Our numerical results also show that user association schemes which favor associating users with pico BSs without taking load balancing into account (e.g. "Picocell First" and "Range Extension"), do not necessarily perform very well.

Chapter 5

A Comparative Study of The Static and Dynamic Modeling Approaches

5.1 Introduction

There are several ways to model cellular systems. In this thesis, we have focused on two modeling approaches, namely *static modeling approach* and *dynamic modeling approach*. We have modeled snapshots of an Hetnet as well as an Hetnet with a dynamic user population via queueing, and we have studied the interplay between user association and resource allocation via these two approaches. Our numerical results show that the engineering insights drawn out of the static study are not always consistent with the insights drawn out of the dynamic study. This can be explained by the fact that these modeling approaches model Hetnets under different

sets of assumptions. In this chapter, we use the numerical results obtained in the previous chapters to compare the static and dynamic modeling approaches, and to draw conclusions on the "robustness" of the results obtained via these two approaches. To do so, we compare trends without any quantitative comparison across the two modeling approaches.

5.2 Comparison of The Resource Allocation Schemes

We compare the three resource allocation schemes (i.e., *PSD*, *OD*, and *CCD*) using the geometric mean rate (*the computed upper bound*) for the static approach, the highest possible arrival rate, the maximum average delay per class, and the average delay in the cell for the dynamic approach. Figures 5.1-5.7 provide the results when the resource allocation parameter K is optimized[1].

Figure 5.1 provides the results for the highest possible arrival rate for the three resource allocation schemes obtained via the dynamic approach. The comparison of the highest possible arrival rate between the three resource allocation schemes shows that *PSD* and *OD* perform significantly better than *CCD* for a large range of K, and that *PSD* performs better than *OD*. The results also show that the system without pico BSs always performs worse than the system with pico BSs. The results in Figures 5.2-5.3, computed via the dynamic approach, show that *PSD*/*OD* and *CCD* perform almost the same (in terms of the maximum average delay per class) for low values

[1]The numerical results show that *PSD* and *OD* have the same delay performance. Because of this, we only show the results for *PSD* in Figures 5.2-5.5.

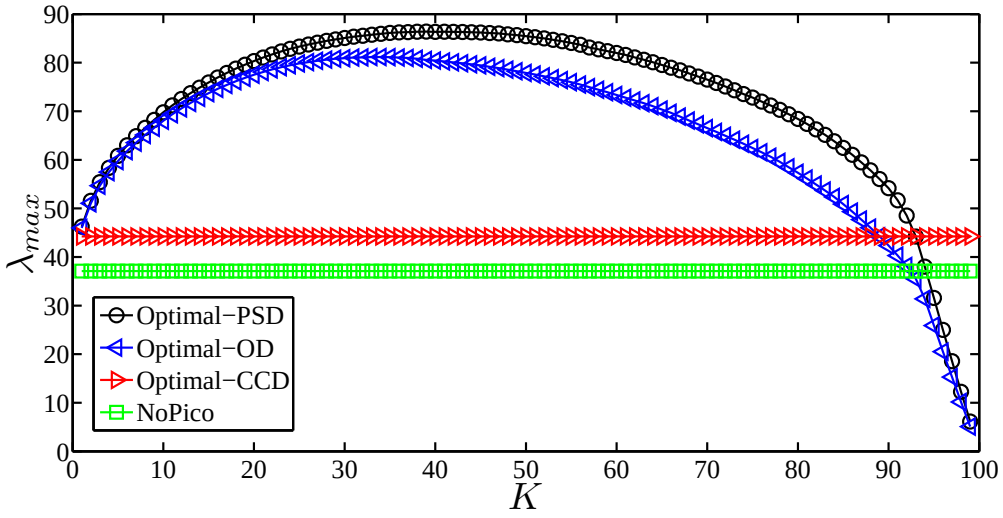


Figure 5.1: PSD, OD, and CCD: The highest user arrival rate as a function of K when $\bar{\rho} = 0.95$, and $F = 10^6$ bits. Note that the results are averaged over 20 networks.

of λ while PSD/OD performs significantly better than CCD for high values of λ. The results also show that PSD/OD and CCD perform significantly better than the system without pico BSs.

Figures 5.4-5.5 provide the results for the average delay in the cell computed via the dynamic approach. The comparison of the average delay in the cell between the resource allocation schemes shows that the system without pico BSs performs better than the system with pico BSs when the arrival rate is quite low. The results also show that for low values of λ, PSD/OD and CCD perform almost the same. However, for high values of λ, PSD/OD performs significantly better than CCD and the system without pico BSs. The results also show that the system without pico BSs always performs worse than the system with pico BSs. Our numerical results show that there is a large difference between PSD/OD and CCD in terms of the highest arrival rate, the maximum average delay per class, and the average delay in the cell, i.e., PSD/OD

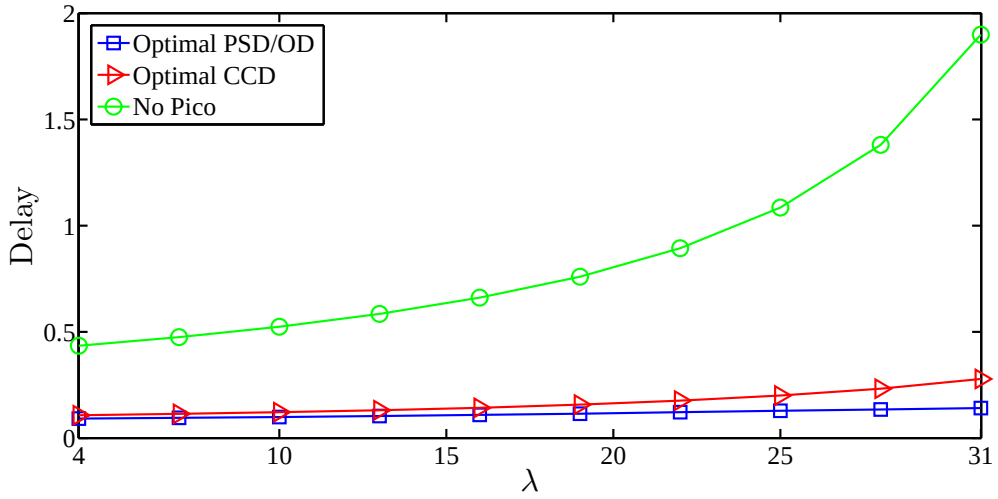


Figure 5.2: PSD, OD, and CCD: The maximum average delay per class as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

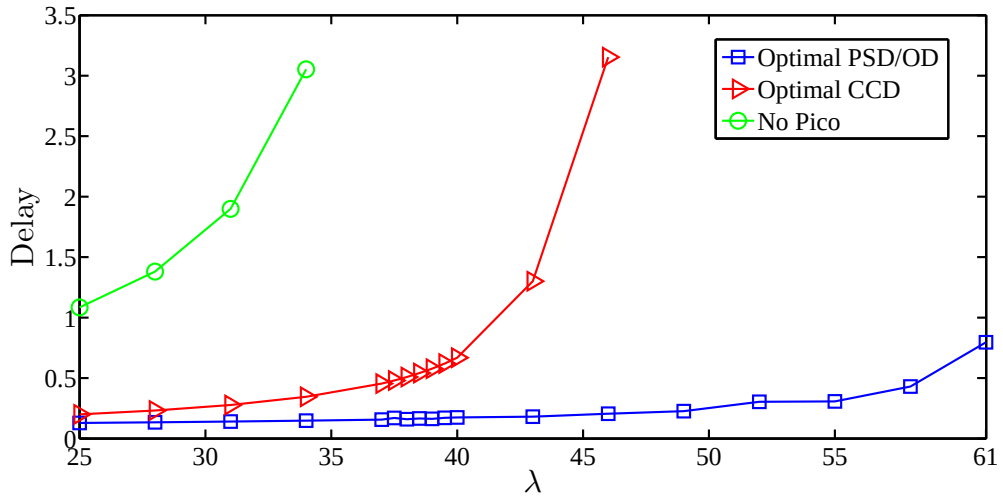


Figure 5.3: PSD, OD, and CCD: The maximum average delay per class as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

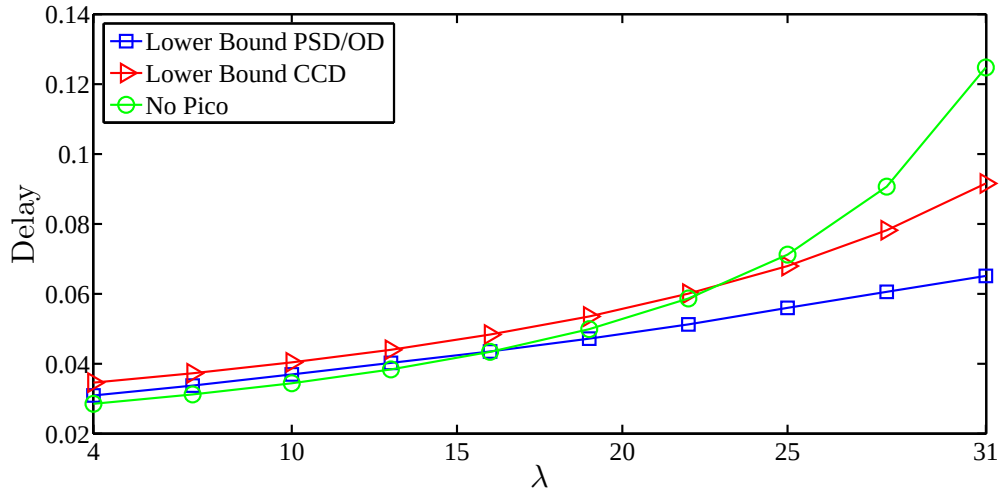


Figure 5.4: PSD, OD, and CCD: The average delay in the cell as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

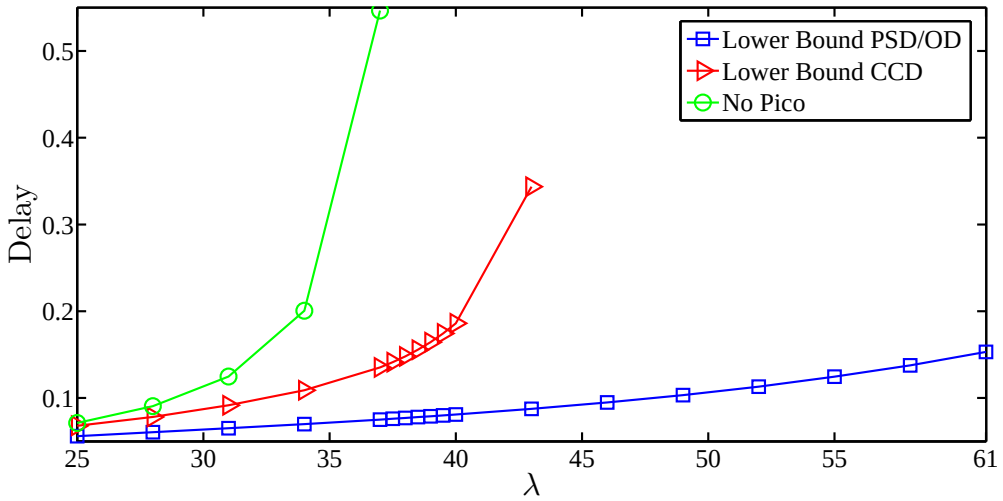


Figure 5.5: PSD, OD, and CCD: The average delay in the cell as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

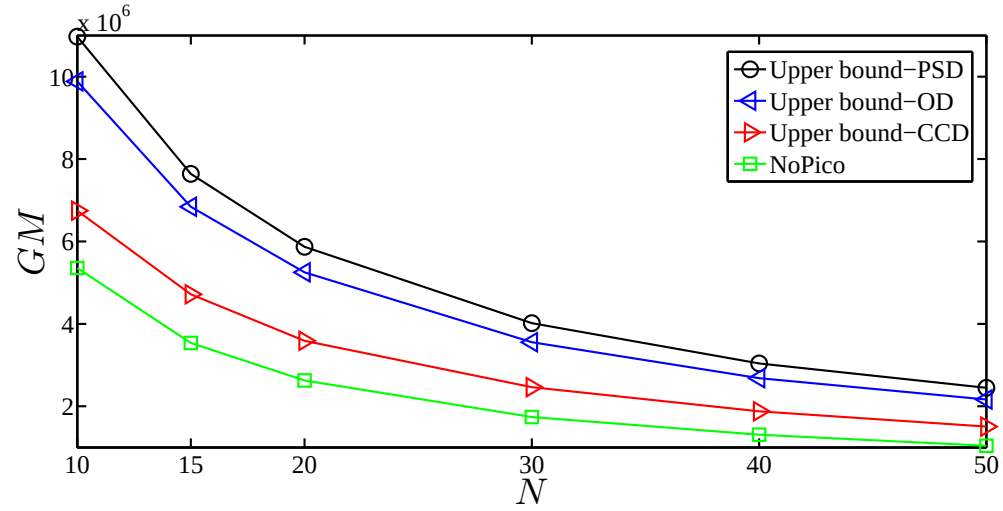


Figure 5.6: PSD, OD, and CCD: Geometric mean rate (in bits per second) as a function of N (the number of users in the cell) when the N users are uniformly distributed in the cell area. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

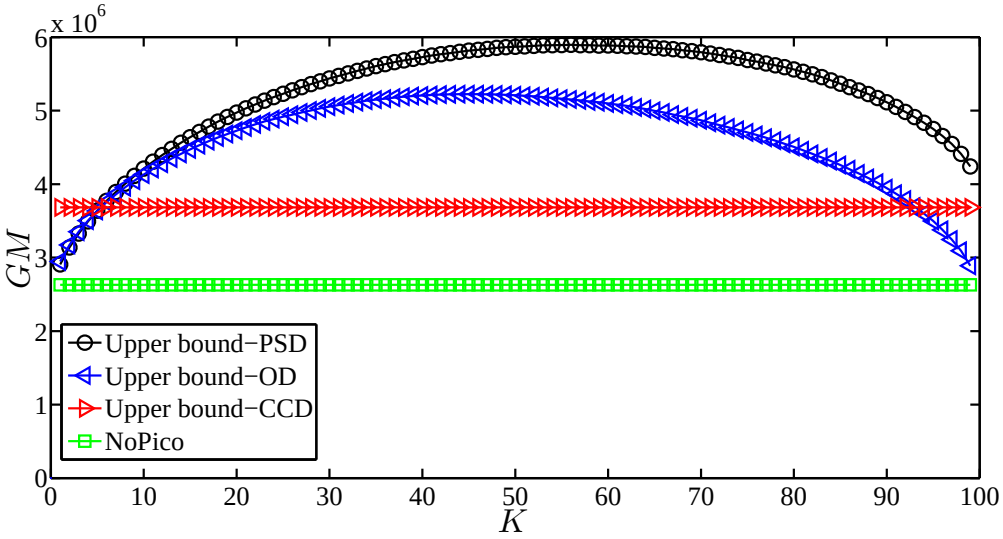


Figure 5.7: PSD, OD, and CCD: Geometric mean rate (in bits per second) as a function of K when $N = 20$ users are uniformly distributed in the cell area. We choose the β that provides the highest geometric mean rate. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

is the best resource allocation for Hetnets with a dynamic user population.

The comparison of the geometric mean rates (obtained via the static approach) between the three resource allocation schemes (using the results in Fig. 5.6) shows that *PSD* and *OD* perform significantly better than *CCD* irrespective of the number of users in the cell N. The results also show that *PSD* performs slightly better than *OD*, and that the system without pico BSs always performs worse than the system with pico BSs. The results in Fig. 5.7 show the geometric mean rate as a function of K when there are 20 users in the cell. The results show that *PSD* and *OD* perform better than *CCD* for a large range of K, and that for all values of K the system with pico BSs performs better than the system without pico BSs.

Our numerical results show that the engineering insights on the RA schemes drawn out the static study are valid in a dynamic context, and vice versa. Note that we have used either an optimal user association or a solution to the relaxed program (either $\mathbf{P'}^{\ell}_{\mathbf{OD}}$ or $\mathbf{P}_{delay}$) to obtain these engineering insights. Next, we focus on *PSD*, and study the impact of user association in more details.

5.3 In Depth Study of PSD

We compare the performance of the association rules *PCF*, *Best SINR*, and *RE* in terms of the geometric mean rate computed via the static approach, the highest possible arrival rate, and the network delay metrics computed via the dynamic approach. The results in Fig. 5.8 show the geometric mean rate of the association rules when the system parameters K and β are optimized. The results show that *PCF* is quasi-optimal since the geometric mean rate of *PCF* is very close to the computed upper

bound (especially when there are more than 20 users in the cell). The results also show that *PCF* is performing better than *Best SINR* and *RE* irrespective of the number of users in the cell (i.e., N), and that the association rules *Best SINR* and *RE* perform almost the same for all possible values of N. Fig. 5.9 provides the results for the geometric mean rate as a function of K when there are 20 users in the system. The results show that all the simple association rules perform better than the system without pico BSs for a large range of K, and that *PCF* with $\beta = 3$ dB outperforms the other user association rules as long as the parameter K is chosen carefully.

In summary, using the *static modeling approach*, we find that *PCF* is performing well, and that the simple association rules *RE* and *Best SINR* are performing almost the same. Our numerical results show that all the simple association rules perform better than the system without pico BSs for a large range of K.

Fig. 5.10 provides the results for the highest possible arrival rate computed via the dynamic approach. The results show that:

- The association rules *PCF* and *Best SINR* perform almost the same, and that these rules perform better than *RE* when K is optimized.

- None of the simple association rules are performing well since none of these rules are close to the optimal user association. The highest arrival rate that can be obtained by the simple UA rules is 62 users per second while the highest possible arrival rate for the *partially shared deployment* is 86 users per second (corresponding to the *optimal* user association). This can be explained by the fact that none of the simple association rules take *load balancing* into account.

- The simple association rules perform better than the system without pico BSs

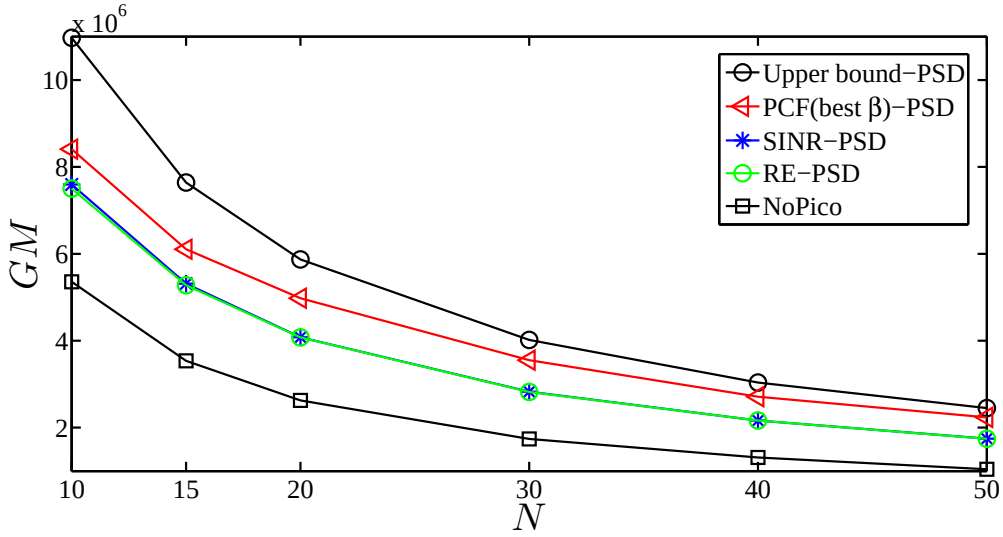


Figure 5.8: PSD: Geometric mean rate (in bits per second) as a function of N (the number of users in the cell) when the N users are uniformly distributed in the cell area. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

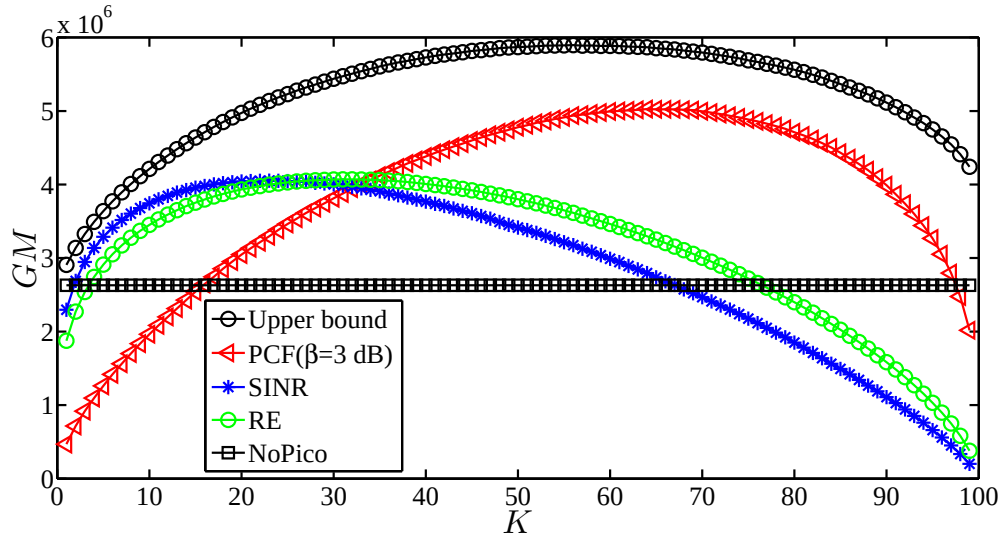


Figure 5.9: PSD: Geometric mean rate (in bits per second) as a function of K when $N = 20$ users are uniformly distributed in the cell area. We choose the β that provides the highest geometric mean rate. Note that the results are averaged over 20 networks, and that for each network, we compute the average results over 100 realizations.

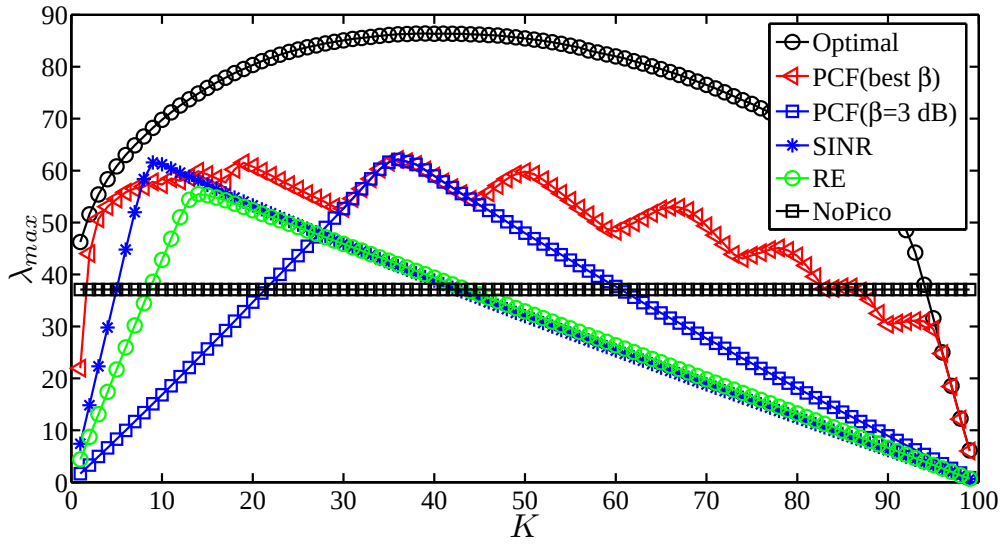


Figure 5.10: PSD: The highest user arrival rate as a function of K when $\bar{\rho} = 0.95$, and $F = 10^6$ bits. For each K, we choose the best β. Note that the results are averaged over 20 networks.

only for a short range K, i.e., when the resource allocation parameter K is not chosen carefully, the system with pico BSs can perform worse than the system without pico BSs

As mentioned above, the system with pico BSs performs worse than the system without pico BSs if the parameter K is not chosen carefully. We now focus on the rules *PCF* and *Best SINR*, and try to understand the impact of K on our network delay metrics. Figures 5.11-5.12 provide the results for the maximum average delay per class as well as the average delay in the cell (computed via the dynamic approach) when K is equal to 10 and 36 for *Best SINR* and *PCF*, respectively. Note that $K = 10$ and $K = 36$ maximize the highest arrival rate for *Best SINR* and *PCF*, respectively. The results show that:

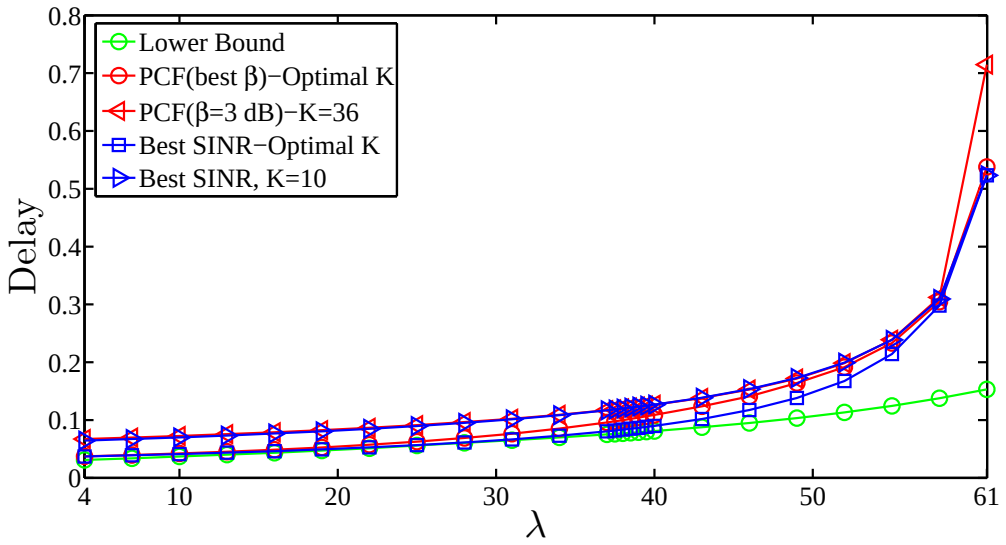


Figure 5.11: PSD, Best SINR, PCF: The average delay in the cell as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

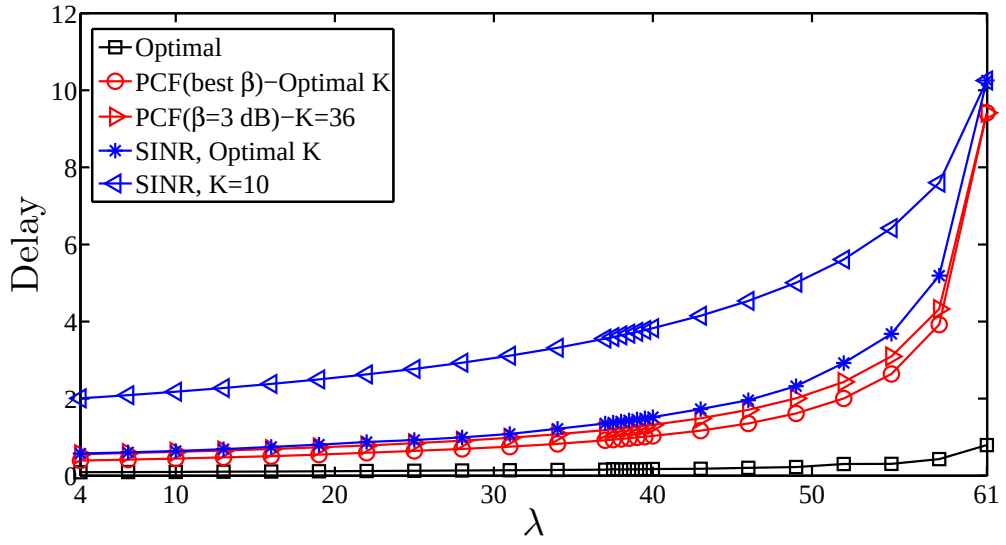


Figure 5.12: PSD, Best SINR, PCF: The maximum average delay per class as a function of λ when $F = 10^6$ bits. Note that the results are averaged over 20 networks.

- There is a negligible difference between the average delay in the cell for *PCF* with $K = 36$ (resp. *Best SINR* with $K = 10$) and the average delay in the cell for *PCF* with the optimal K (resp. *Best SINR* with the optimal K).

- There is a significant difference between the maximum average delay per class for *Best SINR* with $K = 10$ and the maximum average delay per class for *Best SINR* with the optimal K.

Therefore, selecting the value of the parameter K that maximizes the highest arrival rate, does not significantly impact the delay performance of *PCF* while it affects the delay performance of *Best SINR* (especially the maximum average delay per class).

In the *static modeling approach*, *PCF* is quasi-optimal, and the association rules *RE* and *Best SINR* are performing better than the system without pico BSs for a large range of K. However, in the *dynamic modeling approach*, none of the association rules are performing extremely well. The engineering insights drawn out of the dynamic study show that *PCF* and *Best SINR* perform almost the same in terms of the average delay in the cell and the highest possible arrival rate, and that these rules work significantly better than *RE*. These results show that the engineering insights on the association rules in *PSD* drawn out of the static study are not always consistent with the insights drawn out of the dynamic study. Our numerical results also show that the association rule *PCF* is quite robust.

Chapter 6

Conclusion

In this dissertation, we have studied the problem of joint user association and resource allocation in Hetnets that consist of macro and pico BSs. We have considered two modeling approaches, namely a *static modeling approach* and a *dynamic modeling approach*. Using a *static modeling approach*, we have developed a unified framework to study the interplay of user association and resource allocation for Hetnets. In particular, we have formulated joint user association and resource allocation problems. These problems are non-linear integer programs, and hence it is impossible to efficiently obtain optimal solutions. To solve these problems, we have developed techniques to obtain upper bounds on the system's performance. We also have proposed a simple user association rule that performs significantly better than the existing association rules.

Using the proposed framework, we have provided numerical evidence that the proposed techniques compute a tight upper bound on the maximum achievable geometric mean rate. Our numerical results indicate that the resource allocations *par-*

tially shared deployment and *orthogonal deployment* perform significantly better than *co-channel deployment*, and that user association rules which favor associating users with pico BSs yield *significantly better* performance than the conventional association rule.

In the second part of this dissertation, we have studied the problem of joint user association and resource allocation in Hetnets using a *dynamic modeling approach*. We have proposed a unified framework to study the interplay of user association, resource allocation, user arrival, and delay. Given a resource allocation, we have formulated three optimal user association problems to optimize the highest possible arrival rate, the network delay, and the delay-constrained maximum throughput. The proposed problems are non-linear integer programs which are hard to solve efficiently. We have developed numerical techniques to compute either the exact solutions or tight lower bounds to these problems. We have provided numerical evidence that the proposed lower bounds are tight. Our numerical results indicate that *partially shared deployment* and *orthogonal deployment* perform significantly better than *co-channel deployment* if we associate users optimally, and that user association schemes which favor associating users with pico BSs without taking load balancing into account do not necessarily perform very well though they perform better than the other user association rules.

We have used the numerical results obtained in this dissertation to compare the *static and dynamic modeling approaches*, and to draw conclusions on the "robustness" of the results obtained via the two modeling approaches. Our numerical results indicate that the engineering insights on the resource allocation schemes drawn out the static study are valid in a dynamic context, and vice versa. More precisely, the

numerical results obtained in this dissertation show that *partially shared deployment* performs better than *orthogonal deployment*, and that *partially shared deployment* and *orthogonal deployment* perform significantly better than *co-channel deployment* irrespective of the modeling approach.

We have also compared the performance of the simple association rules in *partially shared deployment* via the two modeling approaches. We have provided numerical evidence that the engineering insights on the association rules in *partially shared deployment* drawn out of the static study are not always consistent with the insights drawn out of the dynamic study. The results drawn out of the *static* study indicate that the proposed association rule *Picocell First* performs better than the existing rules, and that it is quasi-optimal. However, the numerical results obtained out of the *dynamic* study indicate that *Picocell First* performs significantly better than the other association rules only for edge users (it does perform as well as the other rules for other users), and that the conventional association rule (i.e., *Best SINR*) performs relatively well except for edge users. Our numerical results, obtained from the dynamic approach, also indicate that user association rules that do not take load balancing into account, do not perform very well in practical cellular systems. The comparative study of the two modeling approaches shows the lack of robustness of certain insights drawn out of the static approach.

Appendix A

Proof of Theorem 1

Let $\mathcal{H}'$ and $\mathcal{H}'_\ell$ denote the set of optimal solutions for $\mathbf{P'_{OD}}$ and $\mathbf{P'^{\ell}_{OD}}$, respectively. The following claim shows that the set of exact solutions to $\mathbf{P'^{\ell}_{OD}}$ is a subset of the set of exact solutions to $\mathbf{P'_{OD}}$ so that solving $\mathbf{P'^{\ell}_{OD}}$ is equivalent to solving $\mathbf{P'_{OD}}$, and vice versa.

Claim 1 *Given problems* $\boldsymbol{P'_{OD}}$ *and* $\boldsymbol{P'^{\ell}_{OD}}$*, we have:*

$$\mathcal{H}' = \left\{ (\{x_{ij}\}, \{\alpha_{ij}\}) \middle| x_{ij}\alpha_{ij} = \frac{x_{ij}}{\sum_{i\in\mathcal{N}} x_{ij}} \right. \\ \left. \{x_{ij}\} \in \mathcal{H}'_\ell,\ \forall i \in \mathcal{N},\ \forall j \in \mathcal{B} \cup \{0\} \right\} .$$

Proof : Let us assume $(\{x_{ij}\}, \{\alpha_{ij}\}) \in \mathcal{H}'$, and there exists some $i_0 \in \mathcal{N}$ and $j_0 \in \mathcal{B} \cup \{0\}$ for which $\alpha_{i_0 j_0} \neq \frac{1}{\sum_{i\in\mathcal{N}} x_{ij_0}}$ while $x_{i_0 j_0} = 1$. For such i's, let us define $\mathcal{U}(j_0) = \{i \in \mathcal{N} \mid x_{ij_0} = 1\}$ If $\alpha'_{ij_0} = \frac{1}{\sum_{i\in\mathcal{N}} x_{ij_0}}$ for all $i \in \mathcal{U}(j_0)$ and $\alpha'_{ij} = \alpha_{ij}$ for all $i \in \mathcal{N}$ and $j \in \mathcal{B} \cup \{0\}$ ($i \notin \mathcal{U}(j_0)$ and $j \neq j_0$), then $(\{x_{ij}\}, \{\alpha'_{ij}\})$ is feasible for $\mathbf{P'_{OD}}$, and according to Lemma 1: $\sum_{i\in\mathcal{U}(j_0)} \log(\lambda'_i) > \sum_{i\in\mathcal{U}(j_0)} \log(\lambda_i)$ where λ'_i

and λ_i are the user i's throughput corresponding to the scheduling coefficients ${\alpha'}_{ij_0}$ and α_{ij_0}, respectively. Hence, there exits another feasible solution $(\{x_{ij}\}, \{{\alpha'}_{ij}\})$ that achieves a larger objective value than $(\{x_{ij}\}, \{\alpha_{ij}\})$. This contradicts the assumption that $(\{x_{ij}\}, \{\alpha_{ij}\}) \in \mathcal{H}'$. The inverse can be proved by using Lemma 1 and following the same argument as above.

Therefore, there exists an onto mapping between the elements of ${\mathcal{H}'}_\ell$ and $\mathcal{H}'$ so that an exact solution $(\{x_{ij}\}, \{\alpha_{ij}\})$ to $\mathbf{P'_{OD}}$ corresponds to an exact solution $(\{x_{ij}\})$ to $\mathbf{P'^{\ell}_{OD}}$ with scheduling coefficients $\frac{x_{ij}}{\sum_{i\in\mathcal{N}} x_{ij}}$, and vice versa. This mapping is not a one-to-one mapping since in some solutions of $\mathbf{P'_{OD}}$ there might exist some $i_0 \in \mathcal{N}$ and $j_0 \in \mathcal{B} \cup \{0\}$ for which $\alpha_{i_0 j_0} > 0$ while $x_{i_0 j_0} = 0$. Note that this does not change users' throughput in $\mathbf{P'_{OD}}$ since $\alpha_{i_0 j_0} x_{i_0 j_0} = 0$. Based on Claim 1 and the structure of problems $\mathbf{P'_{OD}}$ and $\mathbf{P'^{\ell}_{OD}}$, it can be verified that the optimal solutions to $\mathbf{P'_{OD}}$ and $\mathbf{P'^{\ell}_{OD}}$ result in the same users' throughput. Hence, $\mathbf{P'_{OD}}$ and $\mathbf{P'^{\ell}_{OD}}$ are equivalent problems. This completes the proof.

Bibliography

[1] J. G. Andrews, F. Baccelli, and R. K. Ganti. A Tractable Approach to Coverage and Rate in Cellular Networks. *IEEE Trans. on Comm.*, vol. 59, no. 11, Nov. 2011.

[2] S. Sesia, I. Toufik, and M. Baker. *LTE: The UMTS Long Term Evolution: from Theory to Practice*. John Wiley, 2009.

[3] B. H. Walke. *Mobile Radio Networks: Networking, Protocols and Traffic Performance*. John Wiley, 2002.

[4] R1-081957, Categorization of technical proposals for the PHY layer of LTE-A. Qualcomm Europe, 2002.

[5] TR 36.912 V2.0.0, Technical Specification Group Radio Access Network; Feasibility study for Further Advancements for E-UTRA (Release 9). 3GPP, Aug. 2009.

[6] R. Bendlin, V. Chandrasekhar, R. Chen, A. Ekpenyong, and E. Onggosanusi. From Homogeneous to Heterogeneous Networks: A 3GPP Long Term Evolution

Rel. 8/9 Case Study. *in Proc. of the 45th Annual Conference on Information Sciences and Systems (CISS)*, March 2011.

[7] R1-084026, LTE-Advanced Evaluation Methodology. 3GPP, Oct. 2008.

[8] A. Khandekar, N. Bhushan, Ji Tingfang, and V. Vanghi. LTE-Advanced: Heterogeneous Networks. *in Proc. of the 2010 European Wireless Conference (EW)*. pp. 978–982, April 2010.

[9] T. Qu, D. Xiao, and D. Yang. A novel cell selection method in heterogeneous LTE-advanced systems. *in Proc. of Broadband Network and Multimedia Technology*. pp. 510-513, 2010.

[10] 3GPP TS 36.211. Evolved universal terrestial radio access (EUTRA); physical channels and modulation (Realese 8). Technical report, 3GPP TSG R1, Sept. 2007.

[11] H. Claussen. Performance of Macro-and-co-channel Femtocells in a Hierarchical Cell Structure. In *in Proc. of IEEE PIMRC*, 2007.

[12] V. Chandrasekhar and J. G. Andrews. Spectrum Allocation in Tiered Cellular Networks. *IEEE Trans. on Comm.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[13] I. Guvenc, M. Jeong, F. Watanabe, and H. Inamura. A Hybrid Frequency Assignment for Femtocelss and Coverage Area Analysis for Co-Channel Operation. *IEEE Comm. Letters*, vol. 12, no. 12, pp. 880–882, 2008.

[14] J. G. Andrews. Interference Cancellation for Cellular Systems: A Contemporary

Overview. *IEEE Wireless Comm. Magazine*, vol. 12, no. 2, pp. 19–29, April 2005.

[15] M.A. Maddah-Ali, A.S. Motahari, and A.K. Khandani. Communication Over MIMO X Channels: Interference Alignment, Decomposition, and Performance Analysis. *IEEE Trans. on Information Theory*, vol. 54, no. 8, pp. 3457–3470, August 2008.

[16] A. Goldsmith. *Wireless Communications*. Cambridge Univ. Press, 2005.

[17] D. Tse and P. Viswanath. *Fundamentals of Wireless Communications*. Cambridge Univ. Press, 2005.

[18] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic. Interference coordination and cancellation for 4G networks. *in Proc. of IEEE Comm. Magazine*, vol. 47, no. 4, April 2009.

[19] T. Novlan, J. G. Andrews, I. Sohn, R. K. Ganti, and A. Ghosh. Comparison of Fractional Frequency Reuse Approaches in the OFDMA Cellular Downlink. *in Proc. of IEEE GLOBECOM*, Dec 2010.

[20] R. Chang, Z. Tao, J. Zhang, and C.-C. Kuo. A graph approach to dynamic fractional frequency reuse (FFR) in multi-cell OFDMA networks. In *in Proc. of IEEE ICC*, pp. 3993–3998, June 2009.

[21] M. Assad. Optimal fractional frequency reuse (FFR) in multicellular OFDMA system. In *in Proc. of IEEE Vechicular Technology Conference*, pp. 1–5, Sept. 2008.

[22] M. Assad and N. Hassan. Optimal fractional frequency reuse (FFR) and resource allocation in multiuser OFDMA system. In *in Proc. of International Conference on Information and Communication Technologies*, pp. 88–92, Aug 2009.

[23] L. Fang and X. Zhang. Optimal fractional frequency reuse in OFDMA based wireless networks. *in Proc. of the 4th International Conference on Wireless Communications, Networking and Mobile Computing*, Oct. 2008.

[24] Kenya Yonezawa, Kosuke Yamazaki, and Takashi Inoue. Performance Evaluation of Centralized Control Algorithm for Channel Allocation in Pico-cell System. *in Proc. of IEEE Vehicular Technology Conference*, pp. 1659–1663, 2007.

[25] T. Maeyama, T. Inoue, and Y. Takeuchi. A study of channel allocation in Pico-cell system. *Technical Report of IEICE*, pp. 101–105, March 2005. (in Japanese).

[26] K. Yamazaki, K. Yonezawa, H. Ishikawa, and Y. Takeuchi. Performance Evaluation of Novel Frequency Channel Allocation Algorithm for Picocell System. *in Proc. of IEEE PIMRC*, pp. 1659–1663, Sept. 2006.

[27] V. Chandrasekhar and J. G. Andrews. Spectrum Allocation in Tiered Cellular Networks. *IEEE Trans. on Comm.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[28] D. Lopez-Perez, A. Ladanyi, A. Juttner, and Jie Zhang. OFDMA femtocells: A self-organizing approach for frequency assignment. *in Proc. of IEEE PIRMC*, pp. 2202–220, Sept. 2009.

[29] V.Chandrasekhar, J.G. Andrews, Z. Shen, T. Muharemovic, and A. Gatherer.

Power Control in Two-Tier Femtocell Networks. *IEEE Trans. on Wireless Communications*, vol. 8, no. 8, pp. 4316–4328, August 2009.

[30] G. J. Foschini and Z. Miljanic. A simple distributed autonomous power control algorithm and its convergence. *IEEE Trans. on Vehicular Technology*, vol. 42, no. 4, pp. 641–646, Nov. 1993.

[31] A. Damnjanovic, J. Montojo, W. Yongbin, J. Tingfang, L. Tao, M. Vajapeyam, Y. Taesang, S. Osok, and D. Malladi. A Survey on 3GPP Heterogeneous Networks. *IEEE Wireless Comm.*, vol. 18, no. 3, June 2011.

[32] TS 36.300 V10.11.0, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2 (Release 11). 3GPP, Sept. 2013.

[33] M. Cierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting. On Number of Almost Blank Subframes in Heterogeneous Cellular Networks. *in IEEE Trans. on Wireless Comm.*, vol. 12, no. 10, pp. 5061–5073, Oct 2013.

[34] W. Lei, W. Hai, Y. Yinghui, and Z. Fei. Heterogeneous network in LTE-advanced system. *in Proc. of IEEE International Conference on Communication Systems (ICCS)*, 2010.

[35] S. V. Hanly. An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity. *IEEE Journal on Selected Areas in Comm.*, vol. 13, no. 7, Sept. 1995.

[36] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin. A Load-aware handoff and

cell-site selection scheme in multi-cell packet data systems. In *in Proc. of IEEE GLOBECOM*, pp. 3931–3936, 2004.

[37] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin. Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems. *in Proc. of the 10th Annual International Conference on Mobile Computing and Networking (MOBICOM)*, pp. 302–314, 2004.

[38] R. Mathar and M. Schmeink. Integrated optimal cell site selection and frequency allocation for cellular radio networks. *Telecommunication Systems*, vol. 21, no. 2, pp. 339–347, Dec. 2002.

[39] Thierry E. Klein and Seung-Jae Han. Assignment Strategies for Mobile Data Users in Hierarchical Overlay Networks: Performance of Optimal and Adaptive Strategies. *IEEE Journal on Selected Area in Comm.*, vol. 22, no. 5, pp. 849–861, June 2004.

[40] Moon Jung-Min and Cho Dong-Ho. Cell Selection Algorithm Based on Competition of Users in Hierarchical Cellular Networks. *in Proc. of IEEE WCNC*, 2010.

[41] Kemeng Yang, Iqbal Gondal, and Bin Qiu. Multi-Dimensional Adaptive SINR Based Vertical Handoff for Heterogeneous Wireless Networks. *IEEE Comm. Letters*, vol. 12, no. 6, pp. 438–440, June 2008.

[42] NTT DOCOMO. R1-103264, Performance of eICIC with Control Channel Coverage Limitation. 3GPP Std., Montreal, Canada, May 2010.

[43] S. Mukherjee and Ismail Guvenc. Effects of Range Expansion and Interference Coordination on Capacity and Fairness in Heterogeneous Networks. *in Proc. of IEEE Asilomar Conf. on Signals, Systems, and Computers*, pp. 1-5, Nov. 2011.

[44] V. Chandrasekhar, J. G. Andrews, and A. Gatherer. Femtocell Networks: a Survey. *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, Sept. 2008.

[45] H. S. Jo, P. Xia, and J. G. Andrews. Downlink Femtocell Networks: Open or Closed? . *in Proc. of IEEE ICC*, pp. 1–5, June 2011.

[46] K. Son, S. Chong, and G. D. Veciana. Dynamic association for load balancing and interference avoidance in multi-cell networks. *IEEE Trans. Wireless Comm.*, vol. 8, no. 7, pp. 3566–3576, July 2009.

[47] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam. Distributed α-optimal user association and cell load balancing in wireless networks. *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[48] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji. Cell Association and Interference Coordination in Heterogeneous LTE-A Cellular Networks. *IEEE Journal on Selected Areas in Comm.*, vol. 28, no. 9, Dec. 2010.

[49] T. Qu, D. Xiao, D. Yang, W. Jin, and Y. He. Cell selection analysis in outdoor heterogeneous networks. *in Proc. of Advanced Computer Theory and Engineering (ICACTE)*, vol. 5, pp. 554-557, Aug. 2010.

[50] D. Fooladivanda, A. Al Daoud, and C. Rosenberg. Joint Channel Allocation

and User Association for Heterogeneous Wireless Cellular Networks. *in Proc. of PIMRC*, Toronto, Sept. 2011.

[51] D. Lopez-Perez, and X. Chu. Inter-Cell Interference Coordination for Expanded Region Picocells in Heterogeneous Networks. *in Proc. of ICCCN*, 2011.

[52] T. Bu, L. E. Li, and R. Ramjee. Generalized Proportional Fair Scheduling in Third Generation Wireless Data Networks. *in Proc. of INFOCOM*, pp. 1–12, April 2006.

[53] L. Li, M. Pal and Y. R. Yang. Proportional Fairness in Multi-Rate Wireless LANs. *in Proc. of INFOCOM*, pp. 1004–1012, May 2008.

[54] A. Nemirovski. *Interior Point Polynomial Time Methods in Convex Programming.* available at `http://www2.isye.gatech.edu/~nemirovs/Lect_IPM.pdf`.

[55] Qualcomm Incorporated. LTE Advanced: Heterogeneous Networks. White Paper, Feb. 2010.

[56] 3GPP TSG-RAN-WG1, Evolved Universal Terrestrial Radio Access: Further Advancements for E-UTRA Physical Layer Aspects. 3GPP, Tech. Rep. TR 36.814, 2010.

[57] D. Lopez-Perez, A. Ladanyi, A. Juttner, H. Rivano, and J. Zhang. Optimization Method for the Joint Allocation of Modulation Schemes, Coding Rates, Resource Blocks and Power in Self-Organizing LTE Networks. *IEEE INFOCOM*, pp. 111–115, April 2011.

[58] C. Mehlfuhrer, M. Wrulich, J. Ikuno, D. Bosanska and M. Rupp. Simulating the Long Term Evolution physical layer. *in Proc. of the 17th European Signal Processing Conference*, Aug. 2009.

[59] D. Fooladivanda and C. Rosenberg. Joint Resource Allocation and Association for Heterogeneous Cellular Networks. *in IEEE Trans. on Wireless Comm.*, vol. 12, pp. 248-257, Oct 2012.

[60] Y. Qiaoyang, R. Beiyu, C. Yudong, M. Al-Shalash, C. Caramanis, and J.G. Andrews. User Association for Load Balancing in Heterogeneous Cellular Networks. *in IEEE Trans. on Wireless Comm.*, vol. 12, pp. 2706–2716, 2013.

[61] J. Ghimire and C. Rosenberg. Resource Allocation, Transmission Coordination and User Association in Heterogeneous Networks: a Flow-based Unified Approach. *in IEEE Trans. on Wireless Comm.*, vol. 12, pp. 1340–1351, Dec. 2012.

[62] K. Shen and W. Yu. Downlink Cell Association Optimization for Heterogeneous Networks via Dual Coordinate Descent. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[63] Y. Lin and W. Yu. Optimizing User Association and Frequency Reuse for Heterogeneous Network under Stochastic Model. *IEEE Global Communication Conference (Globecom)*, Atlanta, GA, USA, Dec. 2013.

[64] L. Kleinrock. Time-shared systems: a theoretical treatment. *Journal of the ACM (JACM)*, vol. 14, pp. 242–261, April 1967.

[65] L. Kleinrock. *Queueing Systems: Computer applications*. Wiley-Interscience, 1976, ISBN: 978-0471491118.

[66] F. Kelly. *Stochastic Networks and Reversibility*. Wiley, Chichester, 1979.

[67] S.F. Yashkov. Processor sharing queues: Some progress in analysis. *Queueing Systems*, vol. 2, pp. 1–17, June 1987.

[68] M. K. Karray and M. Jovanovic. A Queueing Theoretic Approach to the Dimensioning of Wireless Cellular Networks Serving Variable-Bit-Rate Calls. *IEEE Trans. Vehicular Technology*, vol. 62, no. 6, pp. 2713–2723, July 2013.

[69] F. Glover. Improved linear integer programming formulations of nonlinear integer programs. *Management Science*, vol. 22, no. 4, pp. 455–460, 1975.