Network Management and Decision Making for 5G Heterogeneous Networks

Yifei Huang

A thesis submitted for the degree of

Doctor of Philosophy

Australian National University

Research School of Engineering

College of Engineering and Computer Science

The Australian National University

May 2017

Declaration

The contents of this thesis are the results of original research and have not been submitted for a higher degree to any other university or institution.

The work in this thesis has been published or submitted for publication as journal papers or conference proceedings.

The work in this thesis has been performed while supervised by Dr. Salman Durrani (The Australian National University), Dr. Xiangyun Zhou (The Australian National University), and jointly collaborated with Dr. Ali A. Nasir (King Fhad University of Petroleum and Minerals) and Dr. Pawel Dmochowski (Victoria University of Wellington). The substantial majority of this work was my own.

Yifei Huang
Research School of Engineering,
College of Engineering and Computer Science,
The Australian National University,
Canberra, ACT, 2601,
AUSTRALIA

Education is more than just about learning and a means to obtain knowledge. Education is the most important investment society can make - a tool that crafts our talents into our abilities, and our abilities into our achievements. It is the foundation that helps us better make the decisions that ultimately allow us to become the individuals we were destined to be. My parents have taught me many lessons in my life, and although through my own faults I have forgotten or neglected most of them, the importance of education is one that will forever resonate with me and leave a lasting impression. Thank you, mum and dad, for teaching me this timelessly invaluable lesson.

Acknowledgments

A PhD is ultimately a learning experience, and throughout my journey I have learnt, witnessed and experienced many lessons that will undoubtedly serve me well both professionally and personally. Our experiences and interactions with others shape our personalities and attitudes, so I'd like to acknowledge the following people not just for the direct roles they played during my degree, but also to show my gratitude for the lessons I have learnt and a more profound understanding of the intangible human qualities they have instilled in me.

Firstly, many thanks to my supervisors Salman Durrani and Xiangyun Zhou for teaching me the importance of a professional attitude. Despite having supervised fewer than ten previous PhD students between the two of you, the guidance I have received has led to a new sense of open-mindedness and appreciation for honest work. Salman, your unwavering dedication to bring the best out of your students, and Sean, your insightful technical judgments make you two a superstar supervisory team. I have not only received superior advice, but also been given the freedom to explore my own research avenues, creating a perfect balance of supervision, flexibility, and trust. I will never be your most academically outstanding or prolific student, but in return for your guidance, I hope I have at least helped the two of you evolve as supervisors as much as you have helped me evolve both as a student and as an individual.

Thank you to my collaborators Ali Nasir, Pawel Dmochowski and Howard Yang, and my overseas visit host Jeff Andrews for teaching me the benefits of professional networking. I am honoured to have worked with such fine intellectuals.

Thank you to my colleagues, both past and present, at the ANU communications research group who have shown me that a healthy workplace creates a healthy mind. These people - including, but not limited to, Rod Kennedy, Nan Yang, Shihao Yan, Jing Guo, Biao He, Yirui Cong, Wanchun Liu, Khurram Shahzaad, Abbas

Koohian and Mohammad Shahedul Karim - are the embodiment of ANU's research philosophy. Collectively, you have set the standard for university research groups, because the group has become even greater and more impressive than the sum of its individuals. My only regret is never beating Rod at squash, even despite some embarrassingly ineffective sledging. Also, sorry if I disturbed any of you with stupid questions from time to time, there's only so much the Internet can teach us (for now).

In particular, I'd like to give special honourable mentions to the following three people:

- Noman Akbar - You may be the most multicutural person I have met, so thank you for showing me that culture is no barrier to friendship. We both know that research can take its toll even on the most dedicated of students, so I'm glad you gave me the chance to (occasionally? Or constantly?) be my annoying self, and *still* pretend to laugh at my sarcastic comments. Also, thank you for teaching me that junk food is the best way to get someone's attention.

- Nicole Sawyer - Saying you are my favourite Canberran in the office is like saying sleeping is my favourite daily activity - too easy, because there was never a contest. Thank you for showing me that quality always trumps quantity when it comes to the length of time of friendships. Your friendliness is infectious, your presence joyful, and if there was a dictionary where words could be defined by people and their actions, you would be the definition of ''kind'' (the adjective, not the noun) and all its synonyms. The world needs more people like you.

- Alice Bates - A picture is worth a thousand words, but I believe an emoji is worth a thousand memories. I hope one day there will be an emoji made in your honor so I can use it to relive our memorable experiences together. Your friendship has been a highlight of my three years in Canberra, and I am eternally grateful for your laughs, advice, and insistence that running 14 kilometres can be fun if I train hard enough (it was, because I did). Thank you for teaching me the importance of a work-life balance by being the role model student. Above all, merci d'être l'ami le plus fantastique.

I'd like also to thank the lunch time crew of Ben Ye, Yon Hon Ng, Rajeev Gore, Alex Martin, Yi Zhou and Chuong Nguyen for confirming the most agreed upon and ancient of sayings - lunch time is the most fun part of the day (I may have just made that up, but I dare anyone to refute this). Thank you all also for improving my vocabulary of 9 letter words. If anyone finds out why there is a "z" in "rendezvous", or what the solution to Focus puzzle number 4622 was, please let me know so I can sleep soundly at night.

To all those mentioned here, as Fall Out Boy once sung so enthusiastically, "Thnks fr th Mmrs".

List of Abbreviations

3GPP	3rd Generation Partnership Project
5G	5th Generation
ABS	Almost Blank Subframes
BS	Base Station
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
CoMP	Coordinated Multipoint
CRE	Cell Range Expansion
CSI	Channel State Information
D2D	Device-to-Device
DRx	D2D Receiver
DTx	D2D Transmitter
dB	decibel
dBm	decibel-milliwatts
EB	Exabyte
eICIC	Enhanced Intercell Interference Coordination
FAP	Femto Access Point
FUE	Femto User Equipment
GB	Gigabyte
GHz	Gigahertz
GP	Geometric Program
HetNet	Heterogeneous Networks
Hz	Hertz
IoT	Internet of Things
JFI	Jain's Fairness Index

LOS	Line of Sight
Mb	Megabits
MBS	Macro Base Station
MHz	Megahertz
MIMO	Multiple-Input-Multiple-Output
ms	milliseconds
MUE	Macro User Equipment
NBI	Network Balance Index
PPP	Point Poisson Process
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
UE	User Equipment
ZF	Zero-Forcing

List of Notations

Variable and parameter notations are consistent within each chapter. The following mathematical notations are consistent throughout the entire thesis:

$\mathcal{CN}(\mu, \sigma^2)$	Complex Normal distribution with mean μ and variance σ^2
$\triangleq$	Defined as
$\text{diag}(d_1, d_2, \ldots, d_N)$	Diagonal matrix with diagonal elements $d_1, d_2, \ldots, d_N$
$\|\cdot\|$	Euclidean norm
$\lfloor \cdot \rfloor$	Floor operator
$(\cdot)^T$	Matrix transpose
$(\cdot)^H$	Matrix conjugate transpose
$(\cdot)^+$	Psuedoinverse matrix
∇f	Gradient of f
$\mathbf{I}_n$	Identity matrix of size $n \times x$
$\log_n$	Logarithm of base n
$\det(\cdot)$	Determinant of matrix
$\max(\cdot)$	Maximum operator
$\min(\cdot)$	Minimum operator
$\text{vec}(\cdot)$	Vectorization operator

Abstract

Heterogeneous networks (HetNets) will form an integral part of future cellular communications. With the proper management of network resources and decisions, the coexistence of small cells with macro base stations will improve coverage, data rate and quality of service for users. This thesis investigates critical issues that will arise in HetNets.

The first half of this thesis studies major consequences of the disparity between HetNet tier transmit powers, namely that of interference and load balancing. To reduce the effects of harmful interference to small cell users arising from powerful macro transmissions, we first design a precoding matrix in the form of a generalized inverse, which, unlike conventional precoding methods, allows the base station to target a user specifically to reduce its own interference to that user. Even with a transmit power constraint, the affected user can achieve significant improvement in its interference reduction at the slightly compromise of existing macro users.

Next, we study load balancing by showing the benefits of a dynamic biasing function for cell range expansion over a static bias value. Our findings indicate that a dynamic bias is a more intuitive way to prevent small cell overloading, and that associating closest users first is a preferred association order.

We conclude our study into load balancing by proposing a new notion of network balance. We describe how network balance is different to user fairness, and subsequently define a new metric called the network balance index which measures the deviation of the actual base station load distribution with the expected load distribution. We show using an algorithm that the network balance index is more useful than fairness in improving sum rate for clustered networks.

The second half of this thesis explores more advanced user-centric issues for HetNets. Chapter 5 proposes a user association scheme that achieves high fairness, and considers user association behaviour with network dynamics. In order to reduce the

computation needed to re-associate a large network, we study the probabilities that each user will have to switch associations when a user or base station enters or leaves. In the process, we find that a shrinking network has more effect on user association than a growing one.

Finally, Chapter 6 extends the conventional idea of HetNets to include device-to-device (D2D) communications. We propose a D2D decision making framework that more suitably selects D2D modes for potential D2D pairs by using a two-stage criteria that leads to fewer incorrect D2D mode selections. Once a suitable D2D mode is selected, we demonstrate how to determine optimal or near-optimal power and resource parameters for each mode in order to maximize sum rate. We present a geometric approach to solving the co-channel power control problem, and closed form expressions where possible for orthogonal frequency allocation. Our comprehensive study validates the potential for D2D integration in future cellular communications.

The proposed techniques and insights gained from this thesis aims to illustrate how networks can be better managed and improve their decision making processes in order to successfully serve future users.

List of Publications

The work in this thesis has been published or submitted for publication in the following journals and conferences:

Journals

J1 **Y. Huang**, S. Durrani and X. Zhou, "Interference Suppression using Generalized Inverse Precoder for Downlink Heterogeneous Networks," *IEEE Wireless Communications Letters*, vol. 4, no, 3, pp. 325-328, June 2015.

J2 **Y. Huang**, A. A. Nasir, S. Durrani and X. Zhou, "Mode Selection, Resource Allocation and Power Control for D2D-Enabled Two-Tier Cellular Networks," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3534–3547, Aug. 2016.

J3 **Y. Huang**, S. Durrani, P. Dmochowski and X. Zhou, "A Proposed Network Balance Index for Heterogeneous Networks," *IEEE Wireless Communications Letters*, vol. 6, no. 1, pp. 98-101, Feb. 2017.

Conferences

C1 **Y. Huang**, S. Durrani and X. Zhou, "Interference Nulling for Offloaded Heterogeneous Users Using Macro Generalized Inverse Precoder," *Proc. IEEE ISCIT*, Oct. 2015.

C2 **Y. Huang**, A. A. Nasir, S. Durrani and X. Zhou, "Graphical Generalization of Power Control in Multiuser Interference Channels," *Proc. IEEE AusCTW*, Jan. 2016.

C3 **Y. Huang**, L. Bell, S. Durrani, X. Zhou and N. Yang, "Effects of Load Dependent Dynamic Biasing and Association Order for Cell Range Expansion," *IEEE Proc. ICSPCS*, Dec. 2016.

C4 **Y. Huang**, S. Durrani and X. Zhou, “Base Station Preference Association with Network Dynamics", accepted for publication, *IEEE VTC-Spring*, Sydney, Australia, 2017.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The need to communicate has been a paramount requirement for society throughout history. As societies grew more advanced and widespread, face-to-face communications was no longer sufficient, leading to the development of long distance communications. We have come a long way from the days of light signalling, mail, the telegram, and even fixed line telephony. Nowadays, mobile and cellular communications form a majority of not only digital voice communications, but also video calling and Internet access. The cellular network is now one of the largest and most critical infrastructures, and we now take the services it provides for granted.

Cellular communications began in the early 1980s, where the first generation of mobile phones used analogue signals to connect to the network. By the time 2G came along a decade later, networks transitioned to digital, and since then every decade or so a new generation of technologies has been introduced, improving upon the previous with higher capacities, better coverage, and often new applications. Approaching 2020, the fifth generation, 5G, is on the horizon, and along with the emerging Internet of Things (IoT), people and devices will soon experience the first major step towards truly ubiquitous connectivity.

Conventional cellular networks consists of a base station serving users, whereby users that wish to communicate with each other relay their signals through a network base station. Base station coverage areas are modelled using hexagonal grids [2]. The uplink channel describes data sent from a mobile user to a base station, while the downlink channel describes data sent from the base station to the mobile user. Although easy to study, this model is a simplification of current cellular networks which are much more irregular and complex. Modern demands and us-

age behaviour of mobile users, such as large data rates, high mobility, dense spatial environments and clustered or bursty traffic have made conventional network structures and management methods simply inadequate to cope with our digital society's requirements.

1.1 The Need for HetNets

The expectation and reliance on technology and data, especially those involving wireless connectivity, is growing exponentially. By 2020, the global average subscriber data usage per month will grow to 7 GB per month, or 40 EB per month total [3]. IoT has become a hot topic of discussion in recent times, and estimates put the number of connected devices globally to be as high as 212 billion [4]. With wireless technologies such as Wi-Fi and 4G matured and embraced by many, we are now beginning to witness how IoT is no longer just a buzzword or a hopeful dream, but something achievable or even inevitable in the coming years and decades.

To enable IoT, more devices must connect to each other, and be able to share or transmit more data. As a result, networks no longer must provide higher capacities, but also better coverage, more reliable service, be better orgainsed and make more critical decisions than ever before. The rapid increasing demand for data, coverage and better service, fueled by IoT and an insatiable demand for entertainment, information and convenience means that conventional cellular network structures have become inadequate to support future requirements. Therefore, new paradigms to improve current networks are needed.

Being the most widespread and established wireless communications technology, it is expected 5G will be the most important component for wireless connectivity. Compared to 4G, experts believe that the following are the most important improvements that 5G will bring [5], the first two of which will be focused on in this thesis:

- **Increased data rate:** Up to 1000x peak data rate is expected for 5G, with even cell-edge users potentially achieving 100 Mbps.

- **Improved coverage:** Increased density of the built environment creates cover-

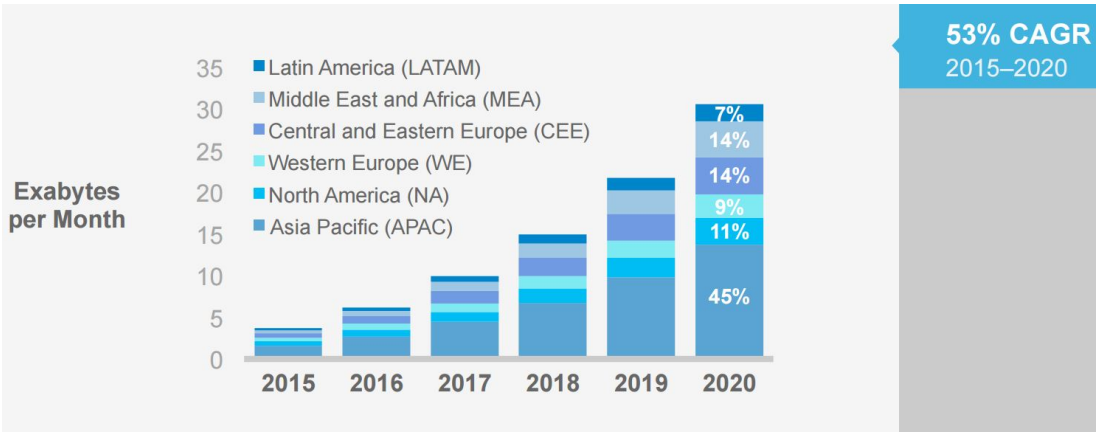

Figure 1.1: Growth of mobile data [1].

age issues for radio signals. For urban areas, dead zones can exist, while one the other hand hotspots can also be overloaded if too many users are accessing the network at once.

- **Reduced latency:** 5G should achieve latency of about 1 ms, which is an order of magnitude faster than current 4G roundtrip times of 15 ms.

- **Improved energy efficiency:** A decrease of 100x in energy efficiency (and by extension cost per bit) is expected. This will also reduce hardware manufacturing costs.

While different researchers and articles in literature have slightly different organisations, there are three key common innovations that will enable the evolution of networks to serve future users and achieve the necessary improvements [5, 6, 7]. The first two, massive multiple-input-multiple-output, or massive MIMO (large number of antennas at a base station), and mmWaves (highly directional, high frequency bandwidth transmission with short ranges), are specific technologies that directly aim at providing increased data rates to users. The third, **heterogeneous networks (HetNets)**, is an evolutionary shift in network structure, and one that encompasses a multitude of technologies, and along with it, challenges to overcome [8].

In its most basic form, HetNets are wireless networks that consist of both macro base stations and small cells, which are its most distinguishing features [5, 8]. Each

Figure 1.2: HetNets will consist of macro base stations as well as small cells.

type of cell is collectively known as a **tiers**, with each tier differing in transmit power and coverage area by approximately an order of magnitude:

- **Macrocells:** Macrocells (or macro base stations) are conventional large cellular base stations with transmit powers in the order of 10 Watts and coverage range of kilometres. Macros are the most expensive cellular infrastructure to set up, and can serve the most number of users.

- **Picocells:** Picocells are smaller sized base stations with transmit powers in the order of 1 Watt and a range of 100 metres. Like macro base stations, picocells are operator deployed and managed, but are much easier to set up, requiring no specific towers and can be positioned out of sight, e.g., on sides of or within a building.

- **Femtocells:** Femtocells are similar in size to Wi-Fi routers, and can be easily installed by users for home or office use. Their transmit powers are in the order of 0.1 Watts, and coverage range of 10 metres. Unlike picocells and macro base stations, femtocells connect to the cellular network via an Internet connection. As a result, femtocells can only be supported in places where sufficient Internet bandwidth is available.

In addition, femtocells can be made open, closed, or hybrid access [2]. Open access allows all users within range to connect, while closed access allows only a select number or specific users to connect, and can be a more secure option. Hybrid access allows preferential service to specific users, while still allowing other users to connect. The determination of which type of femtocell access is often a business decision, as femtocell owners may be unwilling to pay for other users to access their resources.

A more comprehensive definition of a HetNet is one that can also include device-to-device communications (D2D), which occur when user devices communicate directly with each other without relaying through a network base station, macro or small cell [9, 10]. While a truly heterogeneous network will also incorporate other radio access technologies such as Wi-Fi, in this thesis we include only small cells and D2D communications in our description and usage of the term "HetNet."

The physical structure of HetNets brings a number of advantages over macro-only networks. In particular, the following benefits are most relevant in supporting future data demand:

- **Increased Density:** Increased densification is the most noticeable change HetNets bring to conventional cellular networks. Small cells overlay the macro network, leading to more users overall that can be supported by the larger number of total base stations.

- **Improved Coverage:** Small cells are predominately used to improve coverage for cellular networks by serving users in hotspots or dead zones. Large macro base stations may not be able to adequately serve dense urban areas with many blockages and non-LOS transmissions, while small cells can be strategically placed to serve users that may be blocked from the macro base station or in areas with high volume traffic.

- **Ease of Deployment:** Small cells are generally much more easily deployed than macro base stations, which require expensive construction and maintenance. Picocells can be placed on the side of buildings and out of sight, while femtocells can be installed easily in a home or office environment, similar to Wi-Fi

routers or other basic plug-and-play devices. This ease of deployment means that networks can be set up much cheaper, providing access to users much faster than constructing a macro base station. Temporary or ad-hoc networks are possible with small cell deployment.

- **Better Service:** Although a small cell's coverage area is smaller than that of a macro base station's, its shorter transmission distance and fewer served users also means that users may generally get a better quality of service. Shorter transmissions distances improve signal strength at a user receiver, while fewer served users means more resources can be distributed to each user.

1.2 Key HetNet Aspects and Challenges

The introduction of small cells and D2D to cellular networks also introduces a number of new technical aspects, as well as exacerbating existing ones. A study into all or even a majority of these would be beyond the scope of any one thesis, and thus we focus on the following key aspects and outline some major challenges for each.

1.2.1 Interference Management

The most obvious issue with base station densification is the introduction of more interference if transmissions occur on the same channel simultaneously, especially on the downlink [11]. Many interference scenarios have been envisioned for HetNets [12], some so severe that any benefits from using small cells may be overwhelmed by strong interference from other base stations. The most severe causes of interference are caused by downlink transmissions since the interference source is a more powerful base station rather than a user device, as is the case on the uplink. Interference to a small cell user from a macro base station is the most widespread and detrimental scenario for HetNets.

To mitigate this interference, scheduling and resource management can be employed where multiple access techniques split time and frequency resources according to each user's needs [13]. For co-channel transmission, power control can be done

to limit the interference while keeping an acceptable level of desired signal, though this is not always possible [14, 15, 16].

More advanced methods require more channel state information (CSI), but have been shown to be effective even without some imperfect CSI [17, 18]. The emergence of MIMO, where transmitters and receivers all have multiple antennas, has allowed signal processing techniques such as beamforming and precoding at the transmitter (usually a macro base station due to its power) that create focused directions for transmission, hence reducing interference and increasing intended signal strength compared to omni-directional or broadcast transmission [19]. These techniques can be classified into linear methods such as zero-forcing (nullifies the effects of the channel) and maximum ratio transmission (pre-multiply signals by the conjugate of the channel, and maximizes the signal strength at the receiver), and non-linear methods such as dirty-paper coding (requires knowledge of interference at the transmitter) [20]. In practice, linear precoding is preferred as the additional improvement in performance of non-linear precoding often does not justify their additional processing and complexity.

Despite the various interference management options available, sometimes it is desirable for an interference source to target a specific user to reduce its interference to, rather than to reduce its overall signal or impact. To do this, more complex precoding structures are needed.

1.2.2 User Association

User association is the process of deciding which base stations should serve which users. Conventional association connects users to the base station from which it receives the maximum average received power [14, 21] However, as will be explained further in the relevant chapters, such a simple association rule can lead to unwanted consequences in a HetNet. Since macro base stations dominate in terms of transmit power, users are more likely to associate with the macro, and as a result a disproportionate number of users will be macro users, leading to an unbalanced and/or unfair network in terms of user rates. Load-aware or quality-of-service aware user

association policies [21, 22] can help mitigate this issue.

The most common goal of user association schemes is to maximize an objective, usually sum rate or a variant of, with various constraints [23, 24]. In the simplest form, this formulation might look like

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & \sum_i \sum_j x_{i,j} \log_2(1+\gamma_{i,j}) \\ \text{subject to} \quad & 0 \leq x_{i,j} \leq 1, \end{aligned} \tag{1.1}$$

where $x_{i,j}$ is the association variable and $\gamma_{i,j}$ is the signal-to-noise-plus-interference (SINR) at user j from base station i. Binary association is when $x_{i,j} = 0$ or 1, while in fractional association $0 \leq x_{i,j} \leq 1$.

Although seemingly straightforward, any notion of optimal association can in fact be computationally difficult and impractical to implement. Therefore, algorithmic approaches have generally been used [24, 25, 26], or novel ones such as game theory [27, 28, 29], matching theory [30], or a combination of these. To further complicate matters, ideal uplink and downlink association may be different, since the best base station to receive from may not be the same as the best base station to transmit to if each base station has vastly different transmission powers or geographical characteristics. Therefore, different approaches may be needed depending the state of the network and users.

An interesting result on user association from an optimization perspective is that even if fractional association is allowed, the optimal association is still very close to binary association [24], indicating that at its heart, user association can be treated as a combinatorial problem.

Being a computationally costly process due to the large amount of information to be known or shared, user association may not need to be performed with every small change in the network. For future networks with high base station and user density, and more complex association schemes, it is worthwhile to study how user associations may change with varying number of users or base stations.

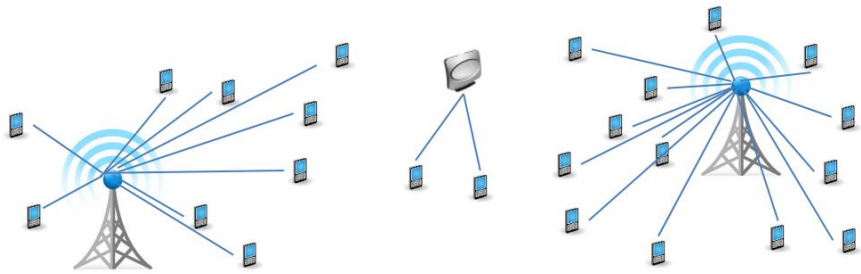

Figure 1.3: Unbalanced network. Pico not utilized fully since only two users are connected.

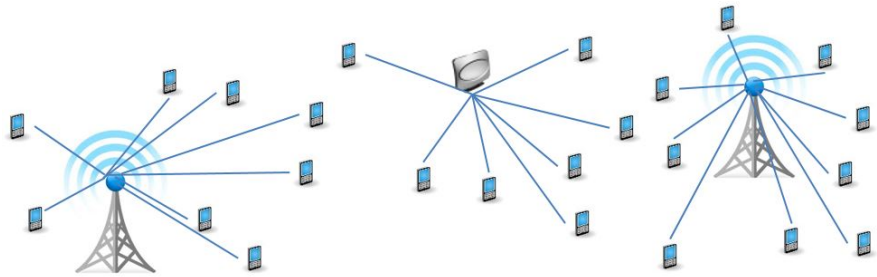

Figure 1.4: More balanced network. Some initial macro users now connect to the pico.

1.2.3 Load Balancing

Related to user association, load balancing is a critical issue further highlighted by HetNets' varying transmission powers compared to conventional macro-only networks. Too many users connected to one base station will overload its limited resources, which must be shared in some manner among all those users (e.g., equal portions if round robin scheduling is used [24, 26]). Since small cells naturally may not pick up as many users as macro base stations, offloading users to small cells is an important procedure to ensure existing users receive adequate service, while also allowing future users options to connect.

To aid offloading and create a more even distribution of users, biasing and cell range expansion (CRE) has been proposed [2, 8, 14] where a bias or weighting is given to a small cell such that users are more likely to associate to a small cell even if its potential data rate may be less than that it would receive from a macro. For instance, if bias is applied to the SINR, a user j would be associated with base station

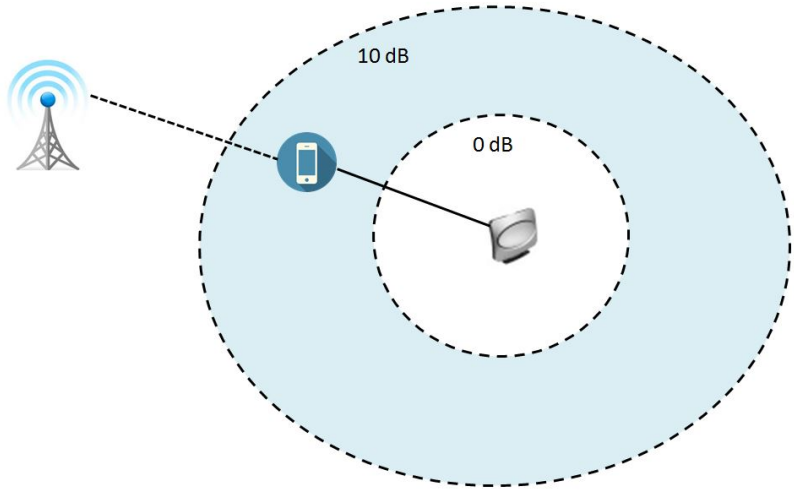

Figure 1.5: A bias value effectively expands the range of small cells (shaded disc region). User previously connected to a macro base station can now connect to a small cell.

i if

$$x_{i,j} = \begin{cases} 1 & \text{if } i = \underset{i}{\operatorname{argmax}} \ \beta_i \gamma_{i,j} \\ 0 & \text{otherwise} \end{cases}, \tag{1.2}$$

where β_i is the bias for base station i. This biasing essentially increases the coverage area or radius of the small cell, and not only relieves pressure off overloaded base stations, but also reduces congestion on backhaul links.

Determining optimal bias values is generally difficult, especially for dense networks, and instead is empirically found through simulations or testing. Typical bias values are in the order of 3-15 dB for picocells [5]. However, optimal bias values may change depending on traffic (i.e., number of users waiting in queues). By definition, macro base stations have no bias, i.e., a bias of 0 dB.

Given that bias values are predetermined and set, it is worth considering the advantages of using dynamic biasing where bias values change depending on the current load on a small cell. For instance, it is intuitive that a small cell would need a smaller bias when it has high load, and a large bias when it has a small load.

On the notion of load balance, there is no consistent definition of "load" in the literature, nor is there one of "balance". Fairness is often used to loosely describe a

balanced network, and the most common quantitative measure of fairness is Jain's Fairness Index (JFI) [31]:

$$JFI = \frac{\sum_{i=1}^{N}(r_i)^2}{N\sum_{i=1}^{N} r_i^2}, \tag{1.3}$$

where r_i is the rate of user i and N is the total number of users. JFI has values in the range $[\frac{1}{N}, 1]$ with JFI = 1 only when all rates are equal.

However, fairness does not represent the same concept as balance, since one can construct a network where a particular base station is overloaded and its users have similar rates, but such a setup is clearly not balanced if nearby base stations have few users. In other words, clustered user distributions tend to be fair, but not balanced. Unlink fairness, there is no accepted quantitative measure of network balance. A study into the differences between fairness and balance for cellular networks does not appear in the literature, despite their similar yet distinct properties.

1.2.4 D2D Communications

D2D communications refers to devices connecting directly to each other without relaying through a base station [10, 32, 33, 34, 35]. This concept could drastically improve many key performance metrics such as data rate and energy efficiency. However, D2D is expected to still be integrated into cellular networks as key decisions would be managed by cellular base stations, rather than D2D pairs exhibiting total autonomy.

To classify D2D resource management, communications can be done either in-band or out-band [36]. In-band refers to D2D using licensed spectrum, while out-band D2D uses unlicensed spectrum. In-band is generally preferred, as out-band usage would also be prone to interference from external sources. For this thesis, all D2D discussions and studies will assume in-band D2D.

The successful integration of D2D communications into the cellular network requires three main procedures:

- **Neighbour Discovery:** Potential D2D pairs must first identify each other and determine certain information such as channel state information [37, 38]. This can be done either autonomously or be network assisted.

- **Mode Selection:** In in-band D2D, there are three modes of operation [39]:

    1. Reuse - Also known as underlay mode. D2D users share existing spectrum resources with other cellular or D2D users.
    2. Dedicated - Also known as overlay mode. D2D users are allocated dedicated spectrum. No interference is present for those D2D pairs.
    3. Cellular - Conventional communications mode. Transmission is relayed through a base station.

    Determining the exact mode for a potential D2D pair is known as mode selection, and is a key process as it dictates the type of resource management.

- **Resource Allocation:** Once a mode is decided, the network must address resource allocation to meet network requirements [39, 40]. For reuse mode, since interference will be an issue at all receivers, suitable transmission powers must be determined. For dedicated and cellular modes (collectively also known as orthogonal modes), maximum transmit powers can be used since orthogonal resources will be allocated. Therefore, time and frequency portion parameters need to be solved for.

The addition of D2D users to the network requires additional network decisions regarding the above procedures. For instance, what is a suitable criteria to determine mode selection? Once this is decided, how should resources be divided, and what are the subsequent resource allocation portions? These challenges are made more difficult if minimum rate requirements are imposed.

1.3 Thesis Outline and Contributions

This thesis proposes techniques and provides insights into the above four key HetNet aspects to enable better network management and decision making. The main body consists of five technical chapters. The first three chapters look into interference management and load balancing, which are key issues that arise out of the differences

in transmit power of HetNet tiers. The second two chapters study extended ideas in HetNets, namely user association and the integration of D2D communications.

For each technical chapter, we first present the key questions that serve as that chapter's motivations and objectives.

- **Chapter 2: Generalized Inverse Precoding for Macro Base Stations**

  *How can a base station **reduce** or **eliminate** its interference to an external user?*

  Chapter 2 formulates the optimal precoder design problem using the generalized inverse structure for suppressing downlink interference from the MBS to a femtocell use equipment (FUE) subject to given power or interference constraints.

  The contributions of the chapter are:

  – We show that a generalized inverse precoder at an MBS can suppress interference to a target FUE without adversely compromising service to current macro user equipments (MUEs). With no base station transmit power constraints, perfect interference nulling to the FUE can be achieved.

  – We then study the case where power and interference constraints are in place. Using Tikhonov regularization, we obtain a closed form relationship between the constraints and the precoder regularization parameter that controls the amount of effort given to interference suppression.

  – Since a Pareto optimal regularization parameter is difficult to determine, and does not allow the MBS to target a specific FUE rate, we present a linear approximation to find a suitable regularization parameter to ensure fairness in the system, which is defined as equal FUE and average MUE user rates. Our results show that a small compromise in the average MUE rate greatly improves the FUE's rate.

  The results of this chapter have appeared in the following publications [41, 42]:

  J1 **Y. Huang**, S. Durrani and X. Zhou, "Interference Suppression using Generalized Inverse Precoder for Downlink Heterogeneous Networks," *IEEE Wireless Communications Letters*, vol. 4, no, 3, pp. 325-328, June 2015.

C1 **Y. Huang**, S. Durrani and X. Zhou, "Interference Nulling for Offloaded Heterogeneous Users Using Macro Generalized Inverse Precoder," *IEEE ISCIT*, Nara, Japan, 2015.

- **Chapter 3: Dynamic Biasing for Cell Range Expansion**

  *What are the effects of using **dynamic biasing** instead of static/constant biasing for cell range expansion?*

  Chapter 3 proposes a more intuitive biasing strategy, namely that of a dynamic bias function which aims to associate more users when a small cell load is low, and uses a small bias to prevent overloading when the load is high.

  The contributions of this chapter are:

  – We propose a load dependent dynamic bias function and study its benefits over a constant bias in the HetNet downlink with and without a pico user quality of service (QoS). Consequently, we investigate different association orders for dynamic biasing, namely outwards-only and inwards-only, and show that inwards-only associates more users, while outwards-only results in larger sum rate and average pico user performance.

  – We derive equivalent static biases and radii for dynamic biasing for both association orders, hence providing an alternative method to empirically determining suitable bias values. Since neither load dependent dynamic bias functions nor association order have received considerable attention in literature in the context of user association [21], we believe that our work provides significant insight into these concepts.

  The results of this chapter have appeared in the following publications [43]:

  C3 **Y. Huang**, L. Bell, S. Durrani, X. Zhou and N. Yang, "Effects of Load Dependent Dynamic Biasing and Association Order for Cell Range Expansion," IEEE ICSPCS, Gold Coast, Australia, 2016.

- **Chapter 4: Network Balance Index**

*How are **user fairness** and **network balance** different, and what is the benefit of considering network balance instead of fairness?*

In order to establish a quantitative measure of balance, we propose a novel network balancing index and show its usefulness compared to fairness in a clustered scenario.

The contributions of this chapter are:

- We next propose a network balance index (NBI) metric that quantifies the deviation of the current load distribution to the expected (i.e., ideally balanced) load distribution, the latter being determined using multiplicatively weighted Voronoi cell areas. While multiplicatively weighted Voronoi cells have been used to analyse the coverage areas of HetNets in [44, 45], their use to describe network balance has not been considered.
- Using a sum rate improvement algorithm that aims to increase NBI, we show how considering balance can be advantageous to considering fairness in a clustered network. We show analytically and via simulations that when users are heavily clustered, increasing the NBI metric also increases the sum rate as underloaded base stations can better serve edge users, while fairness decreases.

The results of this chapter have appeared in the following publication [46]:

J3 **Y. Huang**, S. Durrani, P. Dmochowski and X. Zhou, "A Proposed Network Balance Index for Heterogeneous Networks," *IEEE Wireless Communications Letters*, vol. 6, no. 1, pp. 98-101, Feb. 2017.

- **Chapter 5: Base Station Preference Association and Network Dynamics**

  *Can network states be predicted if there are minor changes? Which users are **mostly likely to change associations**?*

  Chapter 5 proposes a downlink base station preference association rule where users associate with the base station it is ranked highest in. Compared to max received power association, where users determine which base station the user

wants the most, our association associates users to base stations that want it the most.

The contributions of the chapter are:

- We analytically prove this association achieves high JFI by showing that all base stations, regardless of their tier, will tend to associate a similar number of users.
- We study how network dynamics (individual users or base stations entering or exiting the network) affect user associations under this rule. Our analysis provides exact re-association probabilities for users depending on the ranks, and determines the effects of network size, association strength and user preference ranking on re-association probabilities. Distribution of associated ranks is also derived and verified by simulation.
- Our results indicate that there exists a type of user that is most likely to re-associate, and that a shrinking network has more effect on user association than a growing one.

The results of this chapter have been accepted in the following publication:

C4 - **Y. Huang**, S. Durrani and X. Zhou, "Base Station Preference Association with Network Dynamics," accepted for publication, *IEEE VTC-Spring*, Sydney, Australia, 2017.

- **Chapter 6: D2D Mode Selection and Resource Allocation**

  *How might a network **decide** to allow D2D, and if it does, which **mode** and **parameters** should it choose?*

  Chapter 6 proposes a base station assisted *D2D decision making framework* that incorporates mode selection, resource allocation and power control in a two-tier cellular network. The MBS first decides if D2D dedicated mode is permissible or not based on the DTx-DRx separation distance and the availability of orthogonal resources. If not, an interference criteria is used to determine whether the D2D pair should enter reuse mode or remain in cellular mode. Resource

and power allocation is then applied to maximize user sum rates. Compared to joint optimization methods, this multi-stage decision process can arrive at the correct mode and resource allocation in a much more straightforward fashion with less complexity.

The contributions of the chapter are:

- We propose a mode selection method that prioritizes D2D dedicated mode if the D2D pair are close to each other and orthogonal resources are available, and otherwise allows reuse mode if the D2D pair satisfies a strict distance and interference criteria. We show that our proposed decision making framework allows more dedicated D2D users than conventional methods, and allows more correct decisions (i.e., higher rate) when resources are shared.
- For the D2D reuse mode, we (non-trivially) extend the method described in [39] to three dimensions to solve the power allocation problem in a two-tier cellular network. In this process, we first analytically prove that (i) sum SINR is quasi-convex in any number of varying powers and (ii) sum rate has the same derivative behaviour as sum SINR (and hence is almost quasi-convex) when one received power dominates in magnitude over others. Then using these results, we propose a simple approach of finding the corners or vertices of the power region to solve the power allocation problem, which achieves near-optimal performance as compared to exhaustive search.
- For the cellular and D2D dedicated modes, we show that frequency allocation results in higher rates than arbitrary or time sharing resource block allocation. We solve the frequency allocation problem in a two-tier cellular network to maximize the sum rate, while meeting a minimum rate constraint for all the users. We also present general resource allocation methods, where possible, for arbitrary number of users and transmitters.

The results of this chapter have appeared in the following publications [47, 48]:

J2 **Y. Huang**, A. A. Nasir, S. Durrani and X. Zhou, "Mode Selection, Resource Allocation and Power Control for D2D-Enabled Two-Tier Cellular Networks," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3534–3547, Aug. 2016.

C2 **Y. Huang**, A. A. Nasir, S. Durrani and X. Zhou, "Graphical Generalization of Power Control in Multiuser Interference Channels," *Proc. IEEE AusCTW*, Melbourne, Australia, 2016.

Following the technical chapters, Chapter 6 presents overall conclusions and future research directions in the field of HetNets. Two Appendices are also included. The first contains proofs of various theorems and propositions in the technical chapters, while the second details a more general discussion on the geometric approach used to tackle the power control problem described in Chapter 6.

Chapter 2

Generalized Inverse Precoder for Interference Suppression

Key Question: *How can a base station **reduce** or **eliminate** its interference to an external user?*

The successful deployment of HetNets relies on the management of cross-tier interference, e.g., a user equipment (UE) is offloaded to an FAP (which may also occur during cell range expansion) but is suffering from downlink MBS interference [12]. Thus, it is important to investigate solutions which allow an MBS to suppress its interference to HetNet UEs without compromising service to macro UEs (MUEs).

Recently, many papers have focused on interference management in HetNets [11]. Conventional interference management techniques such as resource allocation or scheduling approaches [19] do not allow multiple users to be served simultaneously in shared spectrum environments. Advanced methods such as coordinated multipoint (CoMP) [49], almost blank subframes (ABS) [50], enhanced intercell interference coordination (eICIC) [51], and even cognitive radio based approaches [52, 53] require extensive cooperation and reliable backhaul which may not be always practical in HetNet scenarios. Power control, i.e., increasing the FAP transmit power, can combat cross-tier interference, but in dense networks this will in turn cause significant interference to nearby small cells.

Another possible approach for HetNet interference management is the use of transmit precoding, which involves pre-multiplying signals with a matrix to effectively give a weighting to each signal component for each antenna. For traditional cellular networks, well known methods such as zero-forcing (ZF) [20], regularized

and vector inverses are available [54], but these have disadvantages such as lack of design flexibility (unable to target specific users), signal leakage (regularized inverses do not completely remove interference) and complexity (matrix inverse is costly to compute, especially for large matrices). Interference alignment can also be employed to completely cancel inter-cell interference under certain conditions [55]. However, complete cancellation may not be desirable as the decodability of the interfering signals limits the data rate of the other users [55]. In [56], an approach for cross-tier interference mitigation using precoder codebooks is presented, but is more about precoder *selection* rather than precoder *design*. A precoder design for HetNets is studied in [57], but its focus is energy efficiency rather than interference management. To the best of our knowledge, precoder designs for cross-tier interference management in multi-user downlink HetNet systems have not been presented.

This chapter is organized as follows. We first present the system model and the desired conditions for our precoder. We then show the generalized inverse precoder design under no transmit power constraints, illustrating the ability of our precoder to suppress or eliminate targeted interference. The more practical scenario with a power constraint then follows. Simulation results for both cases are then presented. Finally, we summarize the main findings.

2.1 System Model

Consider an MBS with N antennas serving $N-k$ MUEs and an FAP with N_f antennas serving k FUEs. All MUEs and FUEs employ single antennas. The FUEs are within the MBS cell radius and are receiving interference from the MBS. We assume the FAP has no initial users, but our system can be extended without loss of generality to include any initial users. The system with $k = 1$ is illustrated in Fig. 2.1.

We make the following channel assumptions: (i) the MBS has perfect channel state information (CSI) of its $N-k$ MUEs, (ii) the FAP has perfect CSI of its k FUEs and (iii) the MBS may have imperfect CSI of the MBS-FUE channels due to either imperfect feedback from the FUEs themselves or feedback via a limited backhaul.

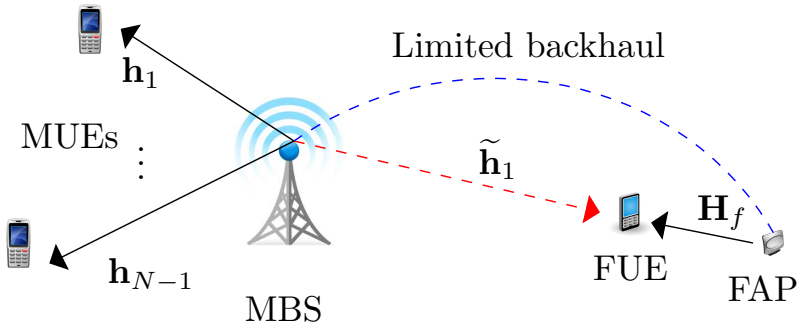

Figure 2.1: System model comprising of MBS, MUEs, FUE and FAP. Interference from MBS to one FUE is illustrated.

Let $\mathbf{H} = \begin{pmatrix} \mathbf{h}_1 & \dots & \mathbf{h}_{N-k} \end{pmatrix}$ denote the channels of the $N-k$ MUEs, where each column $\mathbf{h}_i$ is the $N \times 1$ channel vector for the ith user and each element $\sim \mathcal{CN}(0,1)$. The k MBS-FUEs' channels are similarly defined as $\widetilde{\mathbf{H}} = \begin{pmatrix} \widetilde{\mathbf{h}}_1 & \dots & \widetilde{\mathbf{h}}_k \end{pmatrix}$ whose entries are also $\sim \mathcal{CN}(0,1)$. We denote the MBS precoder as $\mathbf{W}$, which will be defined in the next section.

Taking into account individual MUEs' power allocations, let $\mathbf{Q} = \text{diag}(q_1, q_2 \dots, q_{N-k})$, $q_i \geq 0, \Sigma_{i=1}^{N-k} q_i = N-k$ denote the random MBS power allocation matrix. Using $\mathbf{Q}$, the equivalent precoder is $\mathbf{W} = \mathbf{W}\mathbf{Q}^{\frac{1}{2}}$. Let $\mathbf{x}$ be the vector of $N-k$ independent data streams (one for each MUE) with unit power, and $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2)$ be the independent additive white Gaussian noise (AWGN) vector of dimension $N \times 1$. Thus, using a normalized $\mathbf{W}$ in accordance with our power constraint, the received signals from all MBS transmissions form the vector

$$\mathbf{y} = \sqrt{p_m} \begin{pmatrix} \mathbf{D}^{\frac{1}{2}} \mathbf{H}^H \\ \widetilde{\mathbf{D}}^{\frac{1}{2}} \widetilde{\mathbf{H}}^H \end{pmatrix} \mathbf{W}\mathbf{x} + \mathbf{n}, \tag{2.1}$$

where p_m is the MBS transmit power, and $\mathbf{D} = \text{diag}(\delta_1, \dots, \delta_{N-k})$ and $\widetilde{\mathbf{D}} = \text{diag}(\widetilde{\delta}_1, \dots, \widetilde{\delta}_k)$ are the diagonal pathloss matrices for the MBS-MUE and MBS-FUE channels respectively. The pathloss elements can be determined using well known free-space or industry standard path loss models.

We assume the FAP uses any suitable scheduling scheme to serve its k users during its downlink transmission. Thus, the received signals at the FUEs are denoted

as

$$\mathbf{y}_F = \underbrace{\sqrt{p_f}\mathbf{D}_f^{\frac{1}{2}}\mathbf{H}_F^H\mathbf{x}_F}_{\text{desired}} + \underbrace{\sqrt{p_m}\widetilde{\mathbf{D}}^{\frac{1}{2}}\widetilde{\mathbf{H}}^H\mathbf{W}\mathbf{x}}_{\text{interference}} + \mathbf{n}_F, \tag{2.2}$$

where p_f is the FAP transmit power, $\mathbf{H}_F = \text{diag}(\mathbf{h}_{f,1}, \ldots, \mathbf{h}_{f,k})$ is the $k \times k$ equivalent diagonal Rayleigh fading channel matrix from FAP to FUEs, $\mathbf{D}_f = \text{diag}(\delta_{f,1}, \ldots, \delta_{f,k})$ is the diagonal FAP-FUE pathloss matrix, $\mathbf{x}_F$ is the data $k \times 1$ vector transmitted from the FAP with unit power and $\mathbf{n}_F$ is the $k \times 1$ AWGN vector whose independent elements follow $\sim \mathcal{CN}(0, \sigma^2)$.

The sum rates for the $N - k$ MUEs and any particular FUE are respectively defined as

$$C_{MUE} = \log_2\left(\det(\mathbf{I}_{N-k} + p_m\mathbf{D}\mathbf{H}^H\mathbf{W}\mathbf{W}^H\mathbf{H})\right), \tag{2.3}$$

$$C_{FUE} = \log_2(1 + \gamma), \tag{2.4}$$

where γ is the signal-to-interference-plus-noise ratio at any particular kth FUE with FAP-FUE channel $\mathbf{h}_f$, defined as

$$\gamma = \frac{p_f\delta_{f,k}\left\|\mathbf{h}_{f,k}\right\|^2}{p_m\widetilde{\delta}_k\|\widetilde{\mathbf{h}_k}^H\mathbf{W}\|^2 + \sigma^2}, \tag{2.5}$$

where $\delta_{f,k}$ denotes the pathloss between FAP and kth FUE and $\widetilde{\delta}_k$ denotes the pathloss between MBS and kth FUE.

2.1.1 Problem Statement

Ideally, we desire interference nulling such that at the FUE, SINR = SNR = $\left\|\widetilde{\mathbf{H}}_F^H\right\|$. Any interference nulling technique should not introduce additional inter-user interference for the other $N - k$ MBS users. Thus, we desire a precoding matrix $\mathbf{W}$ which will satisfy the following two conditions:

1. $\mathbf{H}^H\mathbf{W} = \mathbf{I}_{N-k}$. This ensures that the other $N - k$ MBS users still only receive their intended data stream from the MBS, i.e., no additional inter-user interference.

2. Minimize $\left\|\widetilde{\mathbf{H}}^H\mathbf{W}\right\|$. If this can be made to zero, interference from MBS to FUE is nulled. Otherwise, interference is suppressed.

2.2 Precoder Design with No Constraints

In this section we address the precoder design to raise FUEs' rates by suppressing the MBS interference to FUEs, but maintaining interference-free transmission to the MUEs. We design a new precoding matrix for the $N-k$ MUEs using the generalized inverse structure [20]

$$\mathbf{W} = \mathbf{G} + \mathbf{UB}, \tag{2.6}$$

where $\mathbf{G} = (\mathbf{H}^H)^+ = \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}$ is the psuedoinverse of $\mathbf{H}^H$, $\mathbf{U}$ is the $(N-k)\times k$ nullspace of $\mathbf{H}^H$ [58], i.e., $\mathbf{H}^H\mathbf{U} = \mathbf{0}_{N,k}$, and $\mathbf{B}$ is a $k\times(N-k)$ matrix of variable coefficients. The intuition behind using the structure in (2.6) is that the elements of $\mathbf{B}$ can be appropriately chosen to achieve a desired level of interference suppression. In this regard, we will first formulate the problem and then show how to determine $\mathbf{B}$ (and hence the precoder in (2.6)). Finally, we will define a new fairness criteria and show how to determine the precoder accordingly.

2.2.1 Macro Transmission

Consider an MBS with N antennas serving $N-k$ users, with k offloaded users (collectively denoted as UEs) being served by an FAP with N_f antennas. The UEs are still within the MBS cell radius and thus are receiving interference from the MBS. All users employ single antennas. The system with $k=1$ as an example is illustrated in Fig. 1. For simplicity, pathloss and transmit powers are normalized and hence omitted in our formulation as they do not affect the interpretation of our system model. In our simulation results in Section IV, path loss and transmit powers are considered in accordance with 3GPP standards [59].

The channels of the $N-k$ macro users are assumed to be known by the MBS and

can be denoted as

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}_1 & \dots & \mathbf{h}_{N-k} \end{pmatrix}, \tag{2.7}$$

where each column $\mathbf{h}_i$ is the channel vector for the ith user whose elements are independent and identically distributed Rayleigh channels following a complex normal distribution $\sim \mathcal{CN}(0,1)$.

The k offloaded users' channels from the MBS are similarly defined as

$$\widetilde{\mathbf{H}} = \begin{pmatrix} \widetilde{\mathbf{h}}_1 & \dots & \widetilde{\mathbf{h}}_k \end{pmatrix}, \tag{2.8}$$

whose entries follow the same distribution as (1).

If conventional ZF precoding is used [20], let the $N \times (N-k)$ precoder matrix $\mathbf{W}$ for the $N-k$ users be the psuedoinverse of the channel matrix $\mathbf{H}^H$, that is

$$\mathbf{W} = (\mathbf{H}^H)^+ = \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1} = \begin{pmatrix} \mathbf{w}_1 & \dots & \mathbf{w}_{N-k} \end{pmatrix}. \tag{2.9}$$

The received signals by all users due to MBS transmission form the vector

$$\mathbf{y} = \begin{pmatrix} \mathbf{H}^H \\ \widetilde{\mathbf{H}}^H \end{pmatrix} \mathbf{W}\mathbf{x} + \mathbf{n} = \begin{pmatrix} \mathbf{I}_{N-k} \\ \widetilde{\mathbf{H}}^H\mathbf{W} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{N-k} \end{pmatrix} + \mathbf{n},$$

where $\mathbf{x}$ is the vector of $N-k$ independent data streams, one for each macro-served user, and $\mathbf{n} \sim \mathcal{CN}(0,1)$ is the additive white Gaussian noise (AWGN) vector of dimension $N \times 1$ whose independent elements have zero mean and unit power.

2.2.2 Femto Transmission

We assume that FAP has channel knowledge of all its UEs, and uses any suitable transmit scheme to serve them (ZF or scheduling) during its downlink transmission.

Thus, the received signals at the UEs are denoted as

$$\mathbf{y}_F = \underbrace{\mathbf{H}_F^H \mathbf{x}_F}_{\text{desired}} + \underbrace{\widetilde{\mathbf{H}}^H \mathbf{W}\mathbf{x}}_{\text{interference}} + \mathbf{n}_F, \tag{2.10}$$

where $\mathbf{H}_F$ is the $k \times k$ equivalent diagonal Rayleigh fading channel matrix from FAP to UEs, $\mathbf{x}_F$ is the data $k \times 1$ vector transmitted from the FAP and $\mathbf{n}_F$ is the $k \times 1$ AWGN vector whose independent elements follow $\sim \mathcal{CN}(0,1)$. We assume the FAP has no initial users, but our system can be extended without loss of generality to include any initial users.

Assuming the transmit signals $\mathbf{x}_F$ and $\mathbf{x}$ have unit power, the SINR at the kth UE is

$$SINR = \frac{\left\|\mathbf{H}_F^H(k,k)\right\|^2}{\left\|\widetilde{\mathbf{H}}^H \mathbf{W}\right\|^2 + 1}. \tag{2.11}$$

We assume that there is limited feedback between the FAP and MBS rather than full coordination or cooperation. That is, feedback of system features such as SINR is possible, but perfect CSI or transmit data are not exchanged.

2.2.3 Interference Nulling and Suppression

In this section we present a generalized inverse precoder structure which, under perfect CSI conditions, will completely null the MBS interference to the FUE. We also describe three suboptimal but more practical alternative methods of achieving interference suppression if CSI is not perfectly known by the MBS. In all cases, no MBS inter-user interference is introduced.

2.2.3.1 Generalized Inverse Precoder with Perfect CSI

We design a new precoding matrix for the $N-k$ MBS users using the generalized inverse structure

$$\mathbf{W} = (\mathbf{H}^H)^+ + \mathbf{U}\mathbf{B}, \tag{2.12}$$

where $\mathbf{U}$ is the $(N-k) \times k$ nullspace of $\mathbf{H}^H$, i.e., $\mathbf{H}^H\mathbf{U} = \mathbf{0}_{N-k,k}$, and $\mathbf{B}$ is a $k \times (N-k)$ matrix of coefficients. The precoder in (2.12) does not introduce additional inter-user

interference since

$$\mathbf{H}^H\mathbf{W} = \mathbf{H}^H(\mathbf{H}^H)^+ + \mathbf{H}^H\mathbf{U}\mathbf{B} = \mathbf{I}_{N-k}. \tag{2.13}$$

The elements in $\mathbf{B}$ represent the weighting factors for the nullspace vectors of $\mathbf{H}^H$ which can be tuned so as to achieve the desired interference nulling or suppression.

To null interference from MBS to UEs and satisfy condition 2) in 2.1.1, we desire

$$\widetilde{\mathbf{H}}^H\left((\mathbf{H}^H)^+ + \mathbf{U}\mathbf{B}\right) = \mathbf{0}_{k,N-k}, \tag{2.14}$$

which will be only satisfied with an optimal set of coefficients in $\mathbf{B}$ that must be calculated using perfect CSI of the channel $\widetilde{\mathbf{H}}$. If perfect CSI is available at the MBS, the MBS can calculate an optimal $\mathbf{B}$ by expanding brackets in (2.14) and rearranging to obtain

$$\mathbf{B} = -(\widetilde{\mathbf{H}}^H\mathbf{U})^{-1}\widetilde{\mathbf{H}}^H(\mathbf{H}^H)^+. \tag{2.15}$$

For k FAP users, $\widetilde{\mathbf{H}}^H\mathbf{U}$ will be a $k \times k$ square matrix. Thus, $\mathbf{B}$ should always exist as long as $\widetilde{\mathbf{H}}^H\mathbf{U}$ is invertible, which is equivalent to having all channels independent of each other.

Computing $\mathbf{W} = \mathbf{G} + \mathbf{U}\mathbf{B}$ using $\mathbf{B} = -(\mathbf{h}^H\mathbf{U})^{-1}\mathbf{h}^H\mathbf{G}$ involves a psuedoinverse, inverse and nulllspace calculation. We propose a reduced complexity computation method to calculate the same $\mathbf{W}$ in Appendix A.1.

2.2.4 Imperfect CSI

Often, only imperfect CSI may be available at the MBS. We define the imperfect CSI as an erroneous MBS estimate of a particular true MBS-UE channel $\widetilde{\mathbf{h}}$, and is denoted as

$$\widetilde{\mathbf{h}}_{\text{est}} = \widetilde{\mathbf{h}} + \rho\mathbf{e}, \tag{2.16}$$

where $\mathbf{e}$ is random and independent normally distributed error with zero mean and unit variance, and $0 \leq \rho \leq 1$ is a scalar factor representing the degree of imperfection. Substituting $\widetilde{\mathbf{H}}_{\text{est}}$ in place of $\widetilde{\mathbf{H}}$ to calculate 2.15 will give a suboptimal solution and lead to interference suppression.

We note that if imperfect CSI of $\widetilde{\mathbf{h}}$ exists at the MBS, even if the UE continues to be served by the MBS the same ZF precoder $\begin{pmatrix} \mathbf{H}^H \\ \widetilde{\mathbf{h}}_{\text{est}}^H \end{pmatrix}^{-1}$ can still be used and will not affect the other $N-1$ users. This can be summarized as follows:

Proposition 2.1. *Imperfect CSI of one MBS-served user will not affect other users if they have perfect CSI.*

Proof. See Appendix A.2. ■

2.2.4.1 Codebook

In scenarios where the MBS has no access to any CSI, suppose that both FUE and MBS have access to a predetermined orthonormal codebook, such as the Fourier codebook of discrete Fourier transform vectors, e.g.,

$$\text{codebook} = \begin{pmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \dots & \mathbf{f}_N \end{pmatrix}. \tag{2.17}$$

The MBS may use each column as an estimate of $\widetilde{\mathbf{h}}$ to calculate the precoder, and use the vector which yields minimum interference, or the first one which is below an SINR threshold. This method does not require any complex training or adaptive process, and can be used if the MBS has no CSI.

2.2.4.2 Fourier Estimate

In scenarios where reliable channel feedback is not possible between UE and MBS, but CSI is known at the UE due to its detection of MBS pilot signals, UE can exploit the basic feedback capabilities between the FAP and MBS, or simply feedback from UE to MBS if a reliable backhaul doesn't exist. Suppose that $\widetilde{\mathbf{h}}$ is broken down into a linear combination of the codebook basis vectors. That is, $\widetilde{\mathbf{h}}$ can be written as

$$\widetilde{\mathbf{h}} = a_1\mathbf{f}_1 + a_2\mathbf{f}_2 + \ldots + a_N\mathbf{f}_N \tag{2.18}$$

for scalars a_i, $i = 1, \ldots, N$. If the largest two contributors with index values i and j are to be used as channel estimates, the MBS can receive feedback of i, j, a_i, a_j and use

$$\widetilde{\mathbf{h}}_{\text{est}} = a_i \mathbf{f}_i + a_j \mathbf{f}_j. \tag{2.19}$$

In general, the closer the channel estimate to the true channel, i.e., higher order estimate with more codebook components, the greater the interference suppression.

2.3 Precoder Design with Power or Interference Constraint

The previous section solved for a precoder with no power constraints, illustrating that the generalized inverse precoder is a suitable and adaptable method for interference suppression. In reality, base stations have a power constraint, meaning inevitably some loss in performance will be experienced by the MUEs. This section details how to determine precoding parameters to account for power and interference constraints.

2.3.1 Problem Formulation

Using the generalized inverse matrix design, we can formulate the MBS precoder design problem either in terms of an MBS power constraint α as

$$\min \left\| \widetilde{\mathbf{H}}^H (\mathbf{G} + \mathbf{UB}) \right\|^2 \text{ s.t. } \|\mathbf{G} + \mathbf{UB}\|^2 \leq \alpha^2, \tag{2.20}$$

or acceptable MBS-FUE interference limit β as

$$\min \|\mathbf{G} + \mathbf{UB}\|^2 \text{ s.t. } \left\| \widetilde{\mathbf{H}}^H (\mathbf{G} + \mathbf{UB}) \right\|^2 \leq \beta^2, \tag{2.21}$$

which can be shown to be equivalent [60]. Using a least squares approach, (2.20) and (2.21) are also equivalent to

$$\min \left\| \begin{pmatrix} \mathbf{U} \\ \lambda \widetilde{\mathbf{H}}^H \mathbf{U} \end{pmatrix} \mathbf{B} + \begin{pmatrix} \mathbf{G} \\ \lambda \widetilde{\mathbf{H}}^H \mathbf{G} \end{pmatrix} \right\|, \tag{2.22}$$

with the solution

$$\mathbf{B} = -\begin{pmatrix} \mathbf{U} \\ \lambda\widetilde{\mathbf{H}}^H\mathbf{U} \end{pmatrix}^+ \begin{pmatrix} \mathbf{G} \\ \lambda\widetilde{\mathbf{H}}^H\mathbf{G} \end{pmatrix}, \tag{2.23}$$

where $\lambda \geq 0$ is the regularization parameter, which refers to the amount of weighting given to interference suppression, i.e., a larger λ gives more preference to interference suppression. Substituting (2.23) into (2.12), we have

$$\mathbf{W} = \left(\mathbf{I}_N - \mathbf{U} \begin{pmatrix} \mathbf{U} \\ \lambda\widetilde{\mathbf{H}}^H\mathbf{U} \end{pmatrix}^+ \begin{pmatrix} \mathbf{I}_N \\ \lambda\widetilde{\mathbf{H}}^H \end{pmatrix} \right) \mathbf{G}. \tag{2.24}$$

For the MBS to calculate (2.23), $\widetilde{\mathbf{H}}^H$ refers to the MBS estimate of the true MBS-FUE channel. Perfect CSI and $\lambda = \infty$ leads to interference nulling, while imperfect CSI leads to suppression.

2.3.2 Precoder for Given Power or Interference Constraints:

From (2.23) we can see that we need to determine λ in order to find $\mathbf{B}$. Hence, we derive the closed form relationship between λ and the constraints α or β.

Proposition 2.2. *The relationship between λ and α is*

$$\lambda^2 = \frac{1}{\sum_i \mu_i^2} \left(\frac{\sum_i \omega_i}{\alpha - \sum_i \psi_i} - \sum_i \sigma_i^2 \right), \tag{2.25}$$

where σ_i and μ_i are the generalized singular values of $\widetilde{\mathbf{H}}^H\mathbf{U} = \mathbf{L}_1\mathbf{\Sigma}\mathbf{R}^{-1}$ and $\mathbf{U} = \mathbf{L}_2\mathbf{M}\mathbf{R}^{-1}$ respectively, $\mathbf{\Omega} = \mathbf{L}_2\mathbf{M}\left(\mathbf{\Sigma}^H\mathbf{L}_1^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + \mathbf{M}^H\mathbf{L}_2^H(-\mathbf{G})\right)$ and $\mathbf{\Psi} = \mathbf{L}_2^H(-\mathbf{G})$ are both $N \times (N-k)$ matrices, ω_i denotes the elements of $\mathrm{vec}(\mathbf{\Omega})$ and ψ_i denotes the elements of $\mathrm{vec}(\mathbf{\Psi})$. Similarily, the relationship between λ and β is the same as (A.20) but with $\mathbf{\Omega} = \mathbf{L}_1\mathbf{\Sigma}\left(\mathbf{\Sigma}^H\mathbf{L}_1^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + \mathbf{M}^H\mathbf{L}_2^H(-\mathbf{G})\right)$, $\mathbf{\Psi} = \mathbf{L}_1^H(-\widetilde{\mathbf{H}}^H\mathbf{G})$ and replacing α with β.

Proof. See Appendix A.3. ■

Remark 2.1. Proposition 1 gives the value of λ^2 that minimizes the objective for a given power or interference constraint known to the MBS. However, this λ^2 may not be a Pareto optimal value which gives the best 'balance' of objective and constraint.

For example, a predetermined interference constraint may require too much transmit power, so the corresponding λ^2 determined using (A.20) may not be suitable. Calculating such a Pareto optimal λ^2 can be done using methods such as L-curve curvature [60], but these are often very challenging to compute and hence not pursued in this work. Alternatively, the MBS may have a target FUE rate in mind to ensure user fairness. This is investigated below.

2.3.3 Precoder for User Fairness

Suppose we aim to find a λ^2 which achieves user fairness, defined as

$$f(\lambda^2) \triangleq \frac{C_{MUE}}{N-k} - C_{FUE} = 0, \tag{2.26}$$

i.e., when the average MUE rates and FUE rate are equal. Note that other definitions of fairness such as max-min and proportional fair exist in the literature [20]. Max-min fairness is not suitable since the FUE may not initially have the lowest rate, or its rate may not need to be completely maximized at the expense of large MUE rate degradation. Proportional fair is used for scheduling and is also not suitable since it requires past knowledge of user requirements. Hence, a more suitable fairness definition for our setup is when the FUE's rate is equal to the average MUE rates.

A λ^2 that achieves fairness is the root of (2.26), but an exact analytical expression for this is difficult to obtain. Root-finding algorithms can be used to find an approximate solution, but common algorithms such as Newton's method rely on the derivatives of (2.3) and (2.4) with respect to λ^2. Such extensive computation may not be practical, and thus we describe a simpler linear approximation.

The function (2.26) is a logarithmic function, which for some values of its domain exhibits linear behaviour. Through extensive simulations, we have determined that for parameter values in the practical range, the MBS can calculate a suitable λ^2 using a simple linear approximation with respect to λ^2, described in Algorithm 2.1. The initial guesses are arbitrary and chosen empirically, and affect only the approximation error.

Using this linear approximation to calculate the root of $f(\lambda^2)$ ensures that only

Algorithm 2.1 Linear approximation to find root of $f(\lambda^2)$.

Initialize: Initial guesses $\lambda_0^2 = 0$ and $\lambda_1^2 = 2$ (arbitrary)
Approximation: Let $f(\lambda^2) \approx a\lambda^2 + b$
Calculate:
$f(\lambda_0^2) = f(0) = b$
If $b > 0$
  $f(\lambda_1^2) = a\lambda_1^2 + f(0)$
  $a = \frac{f(\lambda_1^2) - f(0)}{\lambda_1^2}$
  Solve: $f(\lambda^2) = 0$
  $\lambda^2 = -\frac{b}{a} = \frac{-\lambda_1^2 f(0)}{f(\lambda_1^2) - f(0)}$
End

three trials are needed. Further, if the first trial using $\lambda^2 = 0$, i.e., $\mathbf{W} = \mathbf{G}$ with no interference suppression, yields a negative value, the FUE is already experiencing better rates than the average MUE, and no additional design is necessary. The approximation error can be reduced if a higher degree polynomial approximation is used, but at the cost of additional computation. However, using a polynomial of degree five or higher may be exceptionally challenging since their roots cannot be found using rational formulas according to the Abel-Ruffini theorem [61].

2.4 Simulation Results

2.4.1 Precoding Without Constraints

We compare the bit error rate (BER) performance with respect to downlink FAP transmission to one offloaded UE of our generalized inverse precoder under interference nulling and suppression scenarios. The conventional ZF precoder ($\mathbf{W} = (\mathbf{H}^H)^+$, i.e., MBS does nothing to reduce its interference to UE) serves as the lower bound for all possible interference suppression methods. For the Fourier estimate, the two largest contributions described by (2.19) are used as estimates, while the codebook method uses all codebook vectors and chooses the one that gives the lowest interference. All simulations use 16-QAM transmission at both the MBS and FAP, with perfect FAP-UE CSI known at the FAP. Precoders are calculated with normalized channel estimates, and interference calculated with normalized precoders.

Table 2.1: Values of Simulation Parameters

Parameter	**Value**
MBS antennas	$N = 8$
FAP antennas	$N_f = 2$
Number of offloaded users	$k = 1$
UE antennas	1
Carrier frequency	2 GHz
Bandwidth	20 MHz
MBS transmit power	43 dBm
FAP transmit power	0 dBm
MBS distance to UE	100 m $\leq d_M \leq$ 500 m
FAP distance to UE	d_F = 10 m
MBS to UE path loss	15.3 + 37.6log$_{10}(d_M)$ (dB)
FAP to UE path loss	38.5 + 20log$_{10}(d_F)$ (dB)
Noise spectral density	-174 dBm/Hz
Imperfect CSI variance	$0.1 \leq \rho \leq 0.5$

Simulation parameters are presented in Table 2.1 [59]. The parameters were chosen so as to reflect worst case values and situations where significant interference is present. The axes of the figures were chosen so as to provide network planning and design insights, such as FAP positioning and hotspot locations.

Fig. 2.2 shows the SINR with varying MBS-UE distance. Even with imperfect CSI of $\rho = 0.5$, an almost 3 dB gain compared to conventional ZF can be made using the generalized inverse precoder, while a 10 dB gain is achieved if $\rho = 0.1$.

Fig. 2.3 plots the BER of the offloaded UE. The FAP transmit power is kept constant while the distance d_M, and hence interference, from the MBS is varied. Interference nulling using perfect CSI results in almost zero BER using the generalized inverse precoder, and thus for figure scaling the plot is not displayed. Comparing the proposed suboptimal methods for interference suppression, using an imperfect CSI estimate for precoder design outperforms the codebook based methods for $\rho = 0.2$, and is far better for even smaller ρ. For instance, for $\rho = 0.1$, an FAP may be placed twice as close to an interfering MBS and still achieve the same BER. Given that such imperfect CSI is often available at the MBS, using these estimates is therefore a practical precoding method. With no CSI knowledge, using a Fourier estimate may be a suitable low complexity alternative. Testing all codebook vectors does outperform

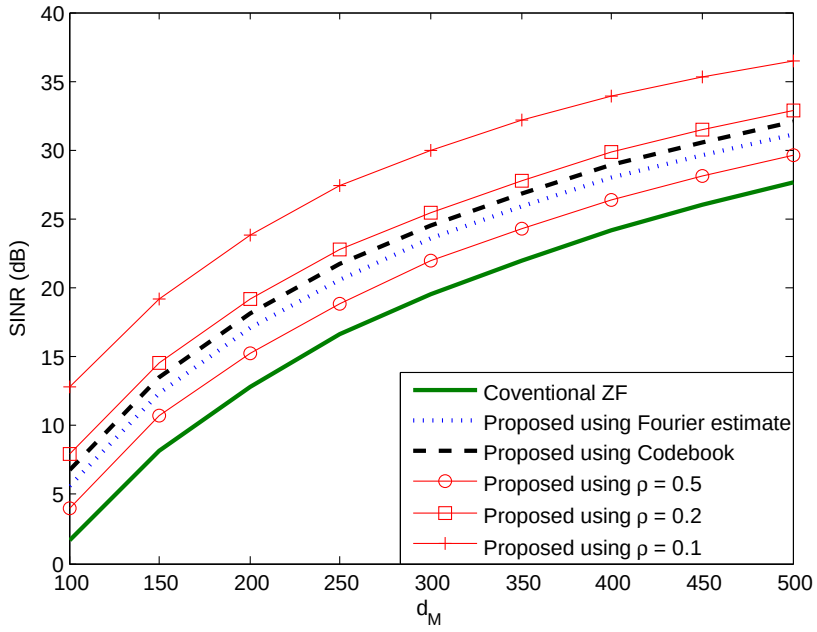

Figure 2.2: MBS to UE distance (d_M) vs SINR at FUE for various interference suppression methods.

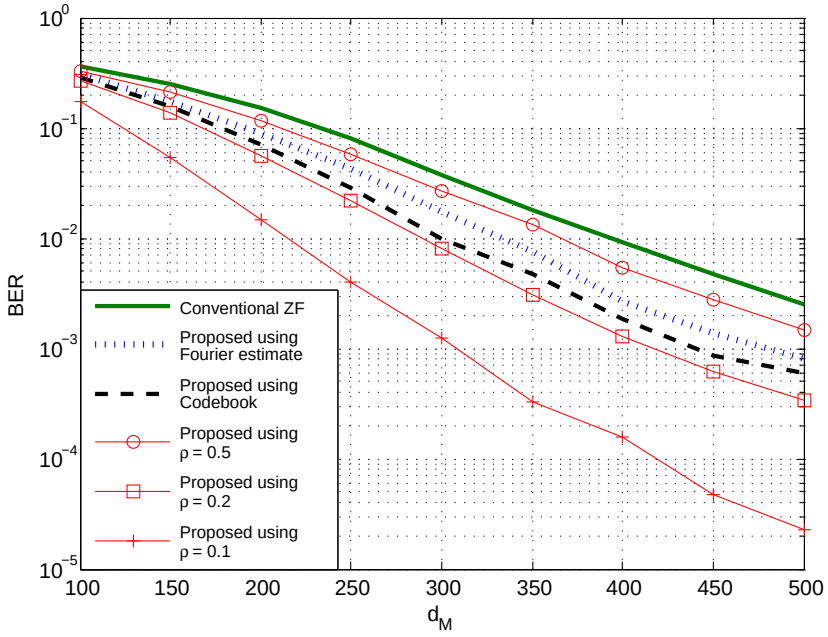

Figure 2.3: MBS to UE distance (d_M) vs BER at FUE for various interference suppression methods.

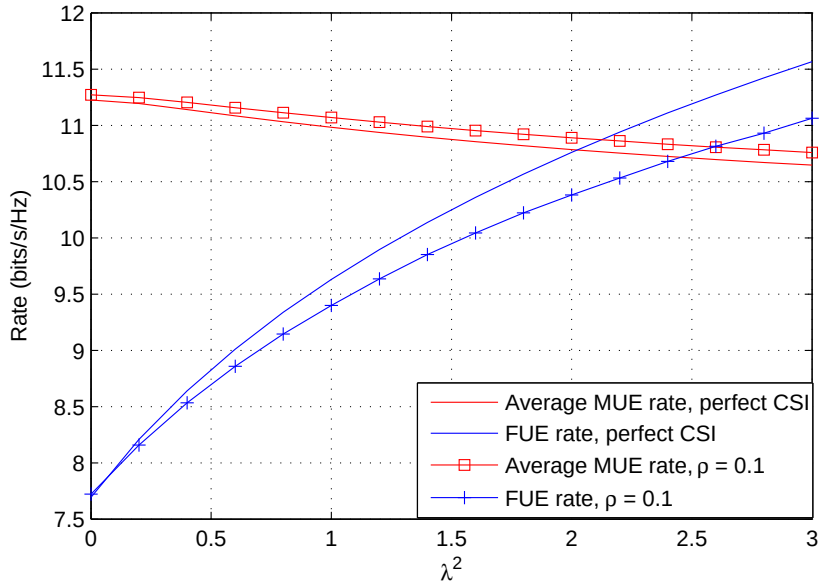

Figure 2.4: Average MUE and FUE rates for with perfect ($\rho = 0$) and imperfect ($\rho = 0.1$) MBS-FUE CSI.

the Fourier estimates but would likely be too computationally extensive for fast fading channels.

2.4.2 Precoding with Power or Interference Constraints

We illustrate the performance of the proposed precoder for user fairness via simulation results averaged over 10,000 Monte Carlo realizations. We consider $N = 8$, $N_f = 2$, $k = 1$ FUE and $N - 1$ MUEs. The MBS-FUE and FAP-FUE distances are fixed at $d_m = 500$ m and $d_f = 10$ m respectively. We set a random MUE power allocation matrix $\mathbf{Q}$, and fix MBS and FAP transmit powers to 20 W and 100 mW respectively. Standard pathloss models are considered for the MBS to FUE and FAP to FUE links, given by 15.3 + 37.6$\log_{10}(d_m)$ dB and 38.5 + 20$\log_{10}(d_f)$ dB [59]. The noise spectral density is set to -174 dBm/Hz and bandwidth is 20 MHz. We also consider the effect of imperfect FUE CSI at the MBS by modelling a particular MBS-FUE channel $\widetilde{\mathbf{h}}$ as $\widetilde{\mathbf{h}}_{\text{est}} = \widetilde{\mathbf{h}} + \rho\mathbf{e}$ where $\mathbf{e} \sim \mathcal{CN}(0, 1)$ is an independent random error vector, and $0 \leq \rho \leq 1$ is the error magnitude.

Fig. 2.4 shows the average MUE and FUE rates for a particular FAP transmit

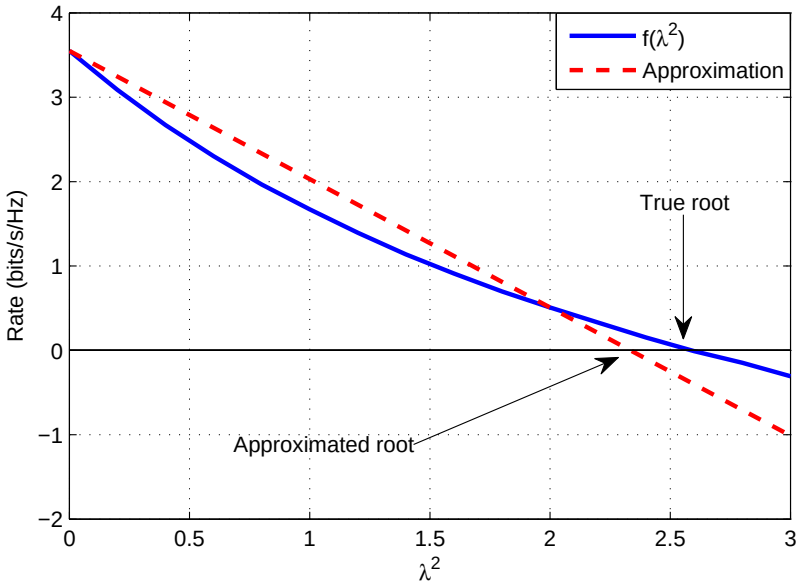

Figure 2.5: Linear approximation of fairness function with $\rho = 0.1$.

power with varying λ^2. The results show the benefit of using the proposed MBS generalized inverse precoder ($\lambda^2 > 0$) compared to conventional ZF with no interference suppression ($\lambda^2 = 0$). We can see that increasing λ^2 gives more preference to interference suppression, and thus improves FUE rate. With an MBS transmit power constraint, i.e., normalized precoding, this compromises the average MUE rate. For our range of simulation parameters, the slopes of the curves indicate that the percentage increase in FUE's rate (>40%) is much greater than the percentage decrease in MUE rates (<5%). Thus, it is evident that significant benefits can be made to the FUE with a small but tolerable decrease in MUE rates. We also observe that more accurate CSI results in a smaller λ^2 to achieve fairness.

Fig. 2.5 illustrates how our linear approximation is used to estimate the root of $f(\lambda^2)$ when $\rho = 0.1$. In this realization, the approximation finds the root to be around $\lambda^2 = 2.3$, while the actual root is around $\lambda^2 = 2.6$. Through extensive simulations, we have observed that the percentage error of the average absolute difference between the rates given by the true fairness λ^2 and the approximated λ^2 is $\approx 5\%$ for practical range of system parameters.

2.5 Summary

We presented in this chapter an MBS precoder structure which, under perfect CSI, can completely null MBS interference to an FUE. Under no transmit power or interference constraints, three practical methods are described for imperfect CSI. Our generalized inverse precoder can achieve significant interference suppression under realistic imperfect CSI values or when using practical suboptimal methods, and will always benefit the offloaded user compared to conventional ZF precoding.

With power or interference constraints, a small compromise for MBS users can drastically improve an FUE's rate. We derive a closed form expression relating the regularization parameter to a given constraint, and present an algorithm to achieve user fairness which requires only three trials. Compared with ZF, the FUE's rate can be significantly improved with a tolerable decrease in the average MUE rate.

Chapter 3

Dynamic Biasing and Association Order for Cell Range Expansion

Key Question: *What are the effects of using **dynamic biasing** instead of static/constant biasing for cell range expansion?*

Currently, HetNet users are associated to the base station (BS) that provides the maximum received signal strength [21]. This may result in the macro being overloaded due to the large disparity in transmit power between base station tiers, leaving small cells under-utilized and overall uneven load distribution [26]. To aid a more even distribution of user association, biasing and cell range expansion (CRE) has been proposed [5, 8] where a bias or weighting is given to a small cell such that users are more likely to associate to a small cell even if its potential data rate may be less than that it would receive from a macro.

Commonly, a constant or static bias value is picked (typical ranging between 3 dB to 15 dB) for a picocell to attract more users [5]. However, each user's service will be affected by the number of other users being served by the same cell on the same subchannel. Therefore, this load dependence means that a constant bias may quickly overload a pico, as the same capturing potential would exist regardless of how much load there already is. In light of this issue, we propose using a dynamic bias as a function of the pico cell load as a natural means to prevent overloading.

The concept of an adaptive/non-constant bias has been teased in the literature, e.g., [62, 63], and shown to be beneficial compared to constant bias value, but falls short of describing a bias function or incorporating load dependence. Optimal bias values are difficult to determine, and are empirically chosen via extensive simula-

tions. In [64], bias values were incremented depending on the ratio of their uplink to downlink demands. An interesting approach was given in [65] that modeled BSs and users as electric charges each with a Gaussian potential function. Notably, it was found that the approach was marginally better than CRE using constant bias values of 3 dB and 6 dB. An optimization approach to user association with load considerations was studied in [24, 66], but a centralized load dependent bias function was not incorporated. Traffic rates as load was studied in [67], although biasing was not used as a means to associate users.

In addition to a changing bias value, the association order is also a critical factor when implementing dynamic CRE. While a constant bias is invariant towards which potential users are associated first, a dynamic bias requires an association order due to the changing effective cell radius. For example, associating farthest possible users first may lead to different results than associating closest users first.

This chapter is organized as follows. We introduce a dynamic bias function with a logistical function expression, then describe two association orders which are necessary to implement dynamic biasing. Equivalent radii and association probability are derived, thereby proving that associating closest users first is more preferable than associating farthest users first. Simluation results comparing dynamic and static biasing verify our hypotheses. Finally, we summarize our main findings.

3.1 System Model

We consider a region with one macro BS and $K-1$ picos, with M users uniformly distributed throughout. Users are associated according to the maximum SINR rule, which is particularly suited to biasing [24]. Users are associated to a BS i if

$$x_{i,j} = \begin{cases} 1 & \text{if } i = \underset{i}{\operatorname{argmax}} \ \beta_i \gamma_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where β_i is the bias value for BS i and $\gamma_{i,j}$ is the SINR at user j from base station i, defined to be

$$\gamma_{i,j} = \frac{P_i |h_{i,j}|^2}{\sum_{k \neq i} P_k |h_{k,j}|^2 + \sigma^2}. \tag{3.2}$$

Here, P_i is the transmit power from BS i, $|h_{i,j}|^2$ is the Rayleigh fading channel gain from BS i to user j with pathloss incorporated (loss exponent denoted by α), and σ^2 is the additive white Gaussian noise power. We assume the macro has no bias.

The rate achieved by user j associated with BS i is load dependent, i.e.,

$$r_{i,j} = \frac{1}{M_i} \log_2(1 + \gamma_{i,j}), \tag{3.3}$$

where M_i is the number of users also served by the same BS on the same subchannel. We have assumed that round robin scheduling is used, due to its proven optimality [24]. Therefore, due to the rate expression's load dependence, user association is a delicate balancing act between associating users to relieve overloading of base stations but also providing the best service. Our primary performance metric is sum rate, given by

$$\sum_{i=1}^{K} \sum_{j=1}^{M} x_{i,j} r_{i,j}. \tag{3.4}$$

For simplicity and investigative purposes, we define load in this chapter as the number of users associated with a BS.

3.2 Dynamic Bias Function

We propose a dynamic bias function that uses a different bias value depending on the number of users already associated with a particular cell. When a cell has few users, a larger bias is used, while if a cell already has a high load, a small or no bias value is used. In essence, dynamic biasing serves as a natural prevention of overloading.

We desire the bias function to decrease slowly with increasing users under low load, then asymptotically approach 0 at high load, resulting in a reverse 'S' or sigmoid shape function. This behaviour has been observed in [63] with the bias value dependent on the traffic arrival rate, and thus we feel is the most suitable shape for

our function. For our work we choose to focus on the logistical function.

3.2.1 Logistical Function

The flipped and shifted logistical function is shown in Fig. 3.1. Three important parameters of the logistical function are A, N_0, and K, where A is the asymptotic maximum value, N_0 is the location of the point of inflection, K is a parameter that controls the steepness of the curve. A dynamic bias function in terms of the number of associated users n can be defined as

$$\beta_i = B(n) = A - \frac{A}{1 + e^{-K(n-N_0)}}. \tag{3.5}$$

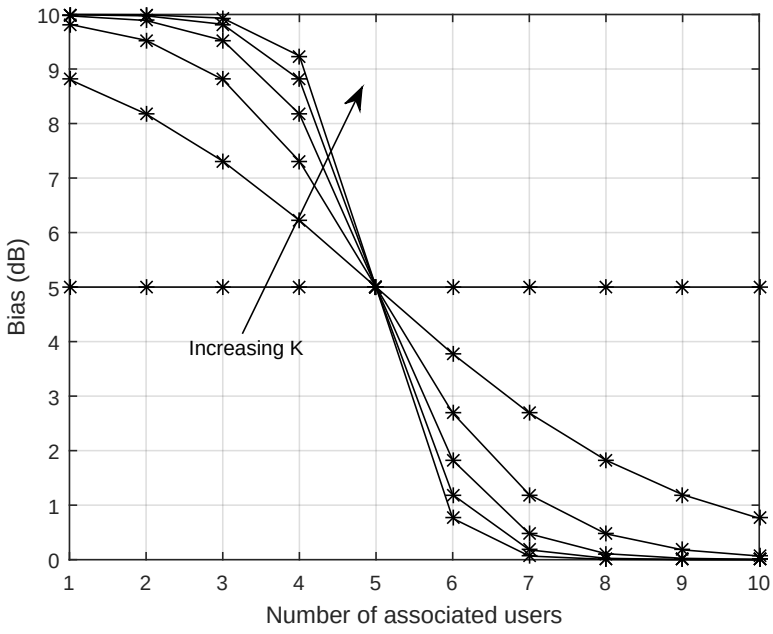

Figure 3.1: Logistical bias function with varying steepness $K = \{0, 0.5, 1, 1.5, 2, 2.5\}$. $A = 10$ dB and $N_0 = 5$ for all functions.

If the parameters A, K and N_0 are chosen suitably, specific functions can be obtained. For instance, setting $K = 0$ will give a constant value at $A/2$, while $K = \infty$ gives a step function with values A and 0.

The integral of the bias function gives an indication of its user capturing potential, and provides a comparative upper bound on the number of users associated.

Fortunately, $B(n)$ has a closed form definite integral:

$$\int_0^N B(n)dn = AN - \frac{A}{K}\ln\left(\frac{\left(1+e^{K(N-N_0)}\right)}{(1+e^{-KN_0})}\right). \tag{3.6}$$

We can confirm that (3.6) $\approx AN/2$ if $N_0 = N/2$, meaning that comparing a dynamic bias function with $N_0 = N/2$ and maximum value A with a constant bias of $A/2$ is a fair comparison as they have similar capturing potential.

3.2.2 Dynamic Biasing and QoS

The study of the potential benefits of dynamic biasing requires consideration of whether a pico QoS, defined to be a minimum rate to be experienced by all pico users, exists, as well as a fair comparison with suitable constant bias values. Due to the shape of the dynamic bias function, a fair comparison with a constant bias would require a logical choice of the constant value. There are two obvious choices for the constant bias value - the maximum dynamic bias value (i.e., A), or the average dynamic bias value (i.e., $A/2$). There are four comparative scenarios:

No QoS, max constant bias: If there are many users to be associated, dynamic biasing would prevent a pico from being overloaded, while the constant bias will keep associating users until all users were associated to a cell. Therefore, dynamic biasing would lead to a higher average pico rate.

No QoS, average constant bias: For dense user deployments, this would have the same benefits as the above scenario but slightly smaller improvements.

QoS, max constant bias: Using a dynamic bias here would associate fewer users than a constant bias, and hence achieve a higher average pico user rate. However, the same effect could be achieved if a lower constant bias is used.

QoS, average constant bias: The same number of users may be associated for both dynamic and constant biasing, but dynamically biased users may have lower average rates as some users may be further away or suffer more interference if they were associated when the bias was high initially.

In light of this, we hypothesize that dynamic biasing is most beneficial if a pico

QoS is not implemented, as a QoS would limit the advantages of a dynamic bias function by directly imposing a limit on the number of associated users and reduce overloading.

3.3 Association Order

With a constant bias and no QoS, the effective cell coverage area also remains constant, and thus users can be associated in any order or at the same time. Since a dynamic bias results in changing effective cell coverage areas, the order of associating users becomes important for loading considerations as some associated users using one order may miss out if a different order is used. We define the following two association orders:

1) Outwards-only: Closest users to a pico are associated first, followed by the next closest, and so on.

2) Inwards-only: Farthest users possible are associated first, followed by the next farthest, and so on.

We hypothesize that generally, with no QoS, inwards-only leads to more users associated with a pico than outwards-only. This will be proved analytically in the coming subsection, but we can illustrate this with an example. From Figs. 3.2 - 3.4, we can see that with increasing load and hence decreasing equivalent radius, users during inwards-only will stay 'ahead' of the shrinking radius until the radius catches up. Those users during outwards-only are meeting towards the radius, leading to the number of users that satisfy the biasing condition running out more quickly.

Importantly, we note that inwards-only generally increases the number of associated users compared to outwards-only, not necessarily increasing the overall sum rate, average pico rate or other metric. For example, an outwards-only association may improve the average pico user rate as associated users would generally be closer to the pico than some of those associated from inwards-only.

If however a QoS is imposed, outwards-only should lead to a better rate performance for both constant and dynamic bias as associated users would be closer to the pico than if inwards-only was used.

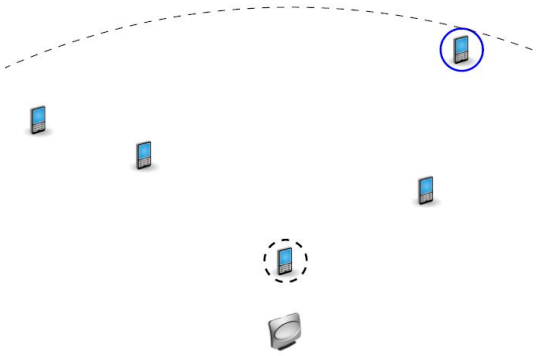

Figure 3.2: Blue users denote those from inwards-only association, dotted outwards-only

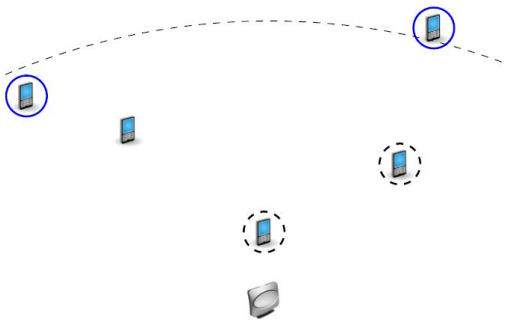

Figure 3.3: Equivalent radius has decreased due to decreasing bias value.

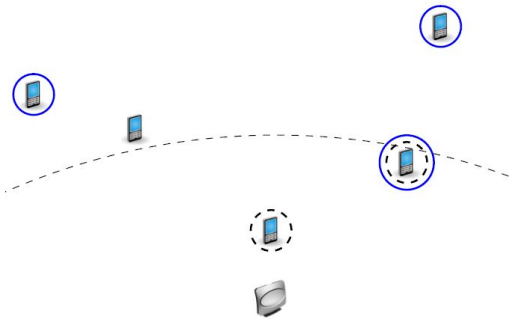

Figure 3.4: Inwards-only associated a 3rd user, but outwards-only associated just 2.

3.3.1 Equivalent Radius

Implementing a dynamic bias will ultimately associate a certain number of users, but this may be alternatively achieved if a specific static bias is used. This also implies that a dynamic bias will result in an equivalent cell radius that corresponds to that static bias. Importantly, we note that these notions of an equivalent bias and radius are approximations and are not perfect implementation substitutes, as they rely on assuming that other base stations are arbitrarily far away.

We can obtain the equivalent cell radius for a given N (number of associated users at which the dynamic bias value is arbitrarily close to 0 dB) and user density λ, and hence show that inwards-only association associates more users than outwards-only. Suppose that a picocell with no biasing has an effective radius of r_0. For outwards-only association, if the last user associated with dynamic biasing (the Nth user) is located outside r_0, this suggests that there must be less than N users initially located within r_0. If the distance of the Nth user to the picocell is r, then the user density is

$$\lambda = \frac{N}{\pi r^2}, \tag{3.7}$$

meaning that the equivalent radius of the dynamic bias is

$$r = \sqrt{\frac{N}{\pi\lambda}}. \tag{3.8}$$

From [63], the equivalent static bias for this radius is then

$$\beta_{out} = \left(\frac{1}{r_0}\sqrt{\frac{N}{\pi\lambda}}\right)^{\alpha} \tag{3.9}$$

where α is the pathloss exponent. For inwards-only, N users outside of r_0 would be associated first before those inside r_0 are associated. Thus, there must be N users located in a disc-region outside of r_0, i.e.

$$\lambda = \frac{N}{\pi(r^2 - r_0^2)}, \tag{3.10}$$

leading to

$$r = \sqrt{\frac{N}{\pi\lambda} + r_0^2}. \tag{3.11}$$

The equivalent bias is

$$\beta_{in} = \left(\frac{1}{r_0}\sqrt{\frac{N}{\pi\lambda} + r_0^2}\right)^{\alpha}. \tag{3.12}$$

By comparing the equivalent radii of outwards-only (3.8) with inwards-only (3.11), we can see that inwards-only will have a larger region of influence, and therefore for the same user density should associate more users.

We observe that the equivalent biases are independent of the exact shape of the bias function, i.e., independent of A and k. This is because the equivalent bias only depends on N, and not the rate at which those N users are associated.

3.3.2 Association Probability

Using the equivalent static biases derived in (3.9) and (3.12), we can also determine the association probability of a typical user associating with a tier-k base station [68]:

$$A_k = \frac{\lambda_k(P_k\beta_k)^{\frac{2}{\alpha}}}{\sum_j^K \lambda_j(P_j\beta_j)^{\frac{2}{\alpha}}} = \left(\sum_{j\neq k}^{K} \lambda_{jk}(P_{jk}\beta_{jk})^{\frac{2}{\alpha}}\right)^{-1}, \tag{3.13}$$

where λ_k is the k-tier base station distribution intensity, $\beta_{jk} \triangleq \beta_j/\beta_k$ and $\lambda_{jk} \triangleq \lambda_j/\lambda_k$. Replacing for β_j, β_k with either (3.9) or (3.12) will give us the association probabilities using outwards-only and inwards-only association respectively.

3.4 Simulation Results and Discussion

We simulate a 1000 m by 1000 m area with 100 users uniformly distributed within. A macro BS is located in the center, with a picocell 250 m away. We model pathloss as d^{-4} where d is the distance between a user and its associated BS[1]. We set the bandwidth to 20 MHz and noise power to -174 dBm/Hz. We assume that the macro and pico share resources and hence interference is present at all users with

[1]The same trends exist for other pathloss exponents, with only absolute values in measured metrics differing.

no interference coordination strategies. Firstly, we simulate scenarios without a pico QoS, then with QoS, imposing a 1 Mb/sec minimum rate for pico users in the latter case.

As there are multiple parameters that can affect the shape of the dynamic bias function, we choose to set $K = 1$ and $N_0 = 10$ while varying the maximum dynamic bias value $A = \{3, 6, 8, 10, 13, 16\}$ dB. We have chosen this range of values for A as the represent typical pico bias values [5]. For a fair comparison, we set the constant bias to be equal to the average bias value of the dynamic bias function, i.e., constant bias is 3 dB lower than the maximum dynamic bias value.

3.4.1 No QoS

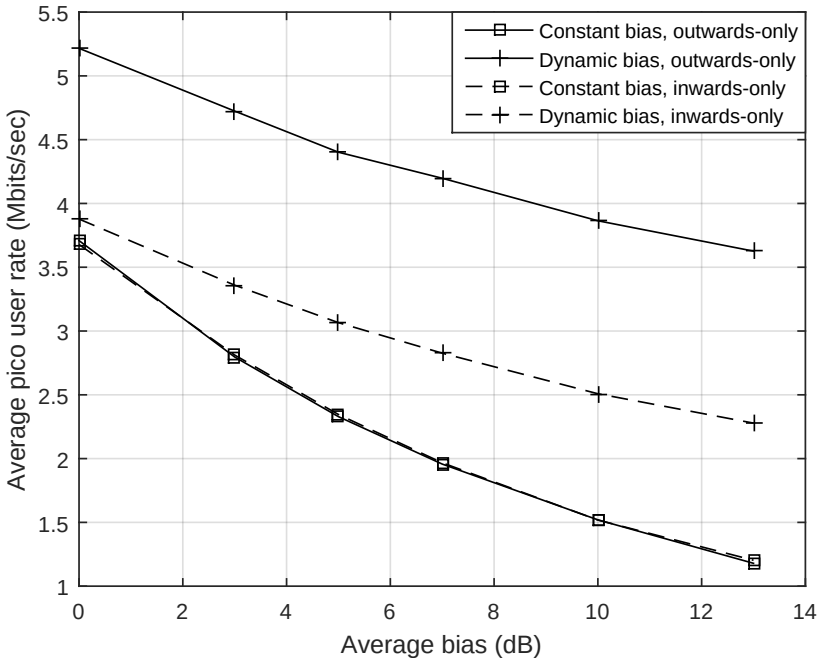

Figure 3.5: Average pico user rate (no pico QoS).

Figs.3.5, 3.6 and 3.7 plot the average bias versus the average pico user rate, pico user association percentage and sum rate respectively with no QoS for pico users. We observe the most benefits of using a dynamic bias over a constant bias when there is no QoS for pico users. Both outwards-only and inwards-only association leads to a higher average pico user rate in Fig. 3.5 than constant biasing as the falling bias

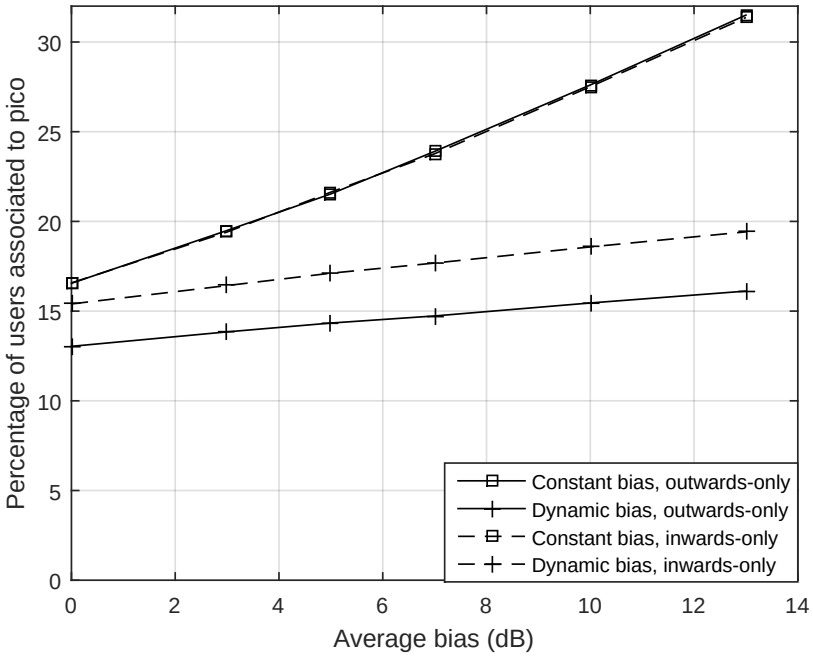

Figure 3.6: Percentage of users associated with the pico (no pico QoS).

values naturally prevents overloading. As expected, constant biasing has the same performance regardless of association order if there is no QoS. The rate at which users are associated with increasing max bias value A is also slower than constant biasing (Fig. 3.6). Further, confirming our analysis from Section 3.3.1, inwards-only associates more users than outwards-only, but has a lower average pico rate.

In terms of sum rate, Fig. 3.7 shows that dynamic biasing with outwards-only results in larger sum rate than constant biasing, but inwards-only gives a lower sum rate. This observation can be explained by the fact that inwards-only may leave users closer to the pico having to associate with the macro if it had already captured too many users, and therefore those users will experience higher interference from the pico.

The performance of constant bias is the same for outwards-only and inwards-only with no pico QoS as all users satisfying the biased association condition in (3.1) will be associated to the pico regardless of which order they are associated in. However, for dynamic biasing, in terms of rate performance outwards-only is more favorable than inwards-only.

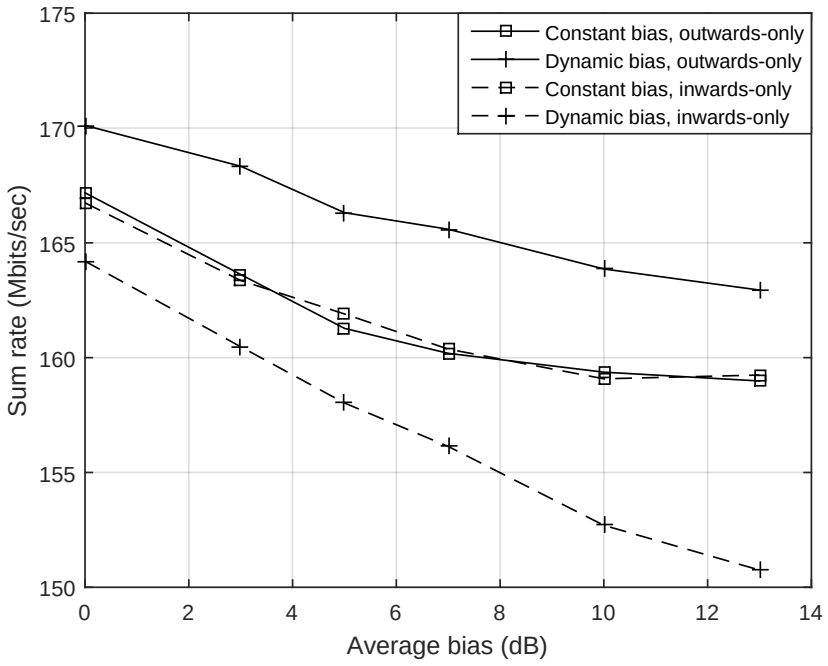

Figure 3.7: Sum rate performance (no pico QoS).

A summary of results with no QoS is provided in Table 3.1.

3.4.2 QoS

If there exists a QoS for pico users, association order affects constant biasing performance also. Figs.3.8, 3.9 and 3.10 plot the average bias versus the average pico user rate, pico user association percentage and sum rate respectively with a QoS for pico users. Dynamic biasing still provides a higher average pico user rate (Fig. 3.8) and associates fewer users (Fig. 3.9) than constant biasing for both outwards-only and

Table 3.1: Dynamic versus constant biasing with no pico user QoS.

No QoS	Outwards-only	Inwards-only
Average pico rate	Dynamic >Constant	Dynamic >Constant
Percentage of users associated to pico	Dynamic <Constant. Rate of increase slower for dynamic than for constant biasing	Dynamic <Constant. Larger percentage than outwards-only
Sum rate	Dynamic >Constant	Dynamic <Constant

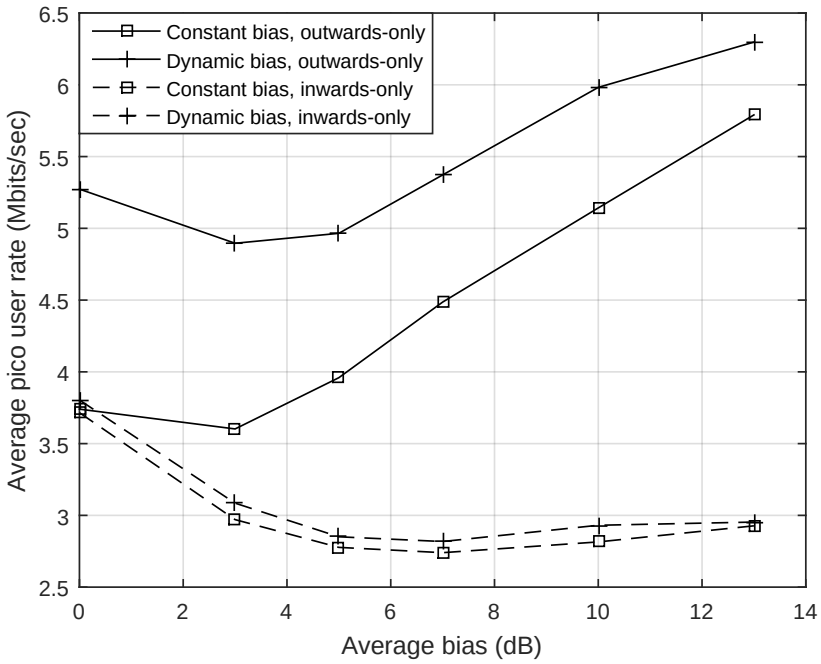

Figure 3.8: Average pico user rate with pico QoS.

inwards-only. Dynamic biasing also only slightly outperforms constant biasing for sum rate (Fig. 3.10) if outwards-only is used, while it performs worse if inwards-only is used. Thus, although we observe mostly the same benefits of dynamic over constant biasing for both association orders, under QoS the improvements are less significant and in fact approach one another for larger bias values.

A summary of results with QoS is provided in Table 3.2.

Table 3.2: Dynamic versus constant biasing with pico user QoS.

QoS	Outwards-only	Inwards-only
Average pico rate	Dynamic >Constant	Dynamic >Constant Improvement less than outwards-only
Percentage of users associated to pico	Dynamic <Constant	Dynamic <Constant.
Sum rate	Dynamic >Constant	Dynamic <Constant

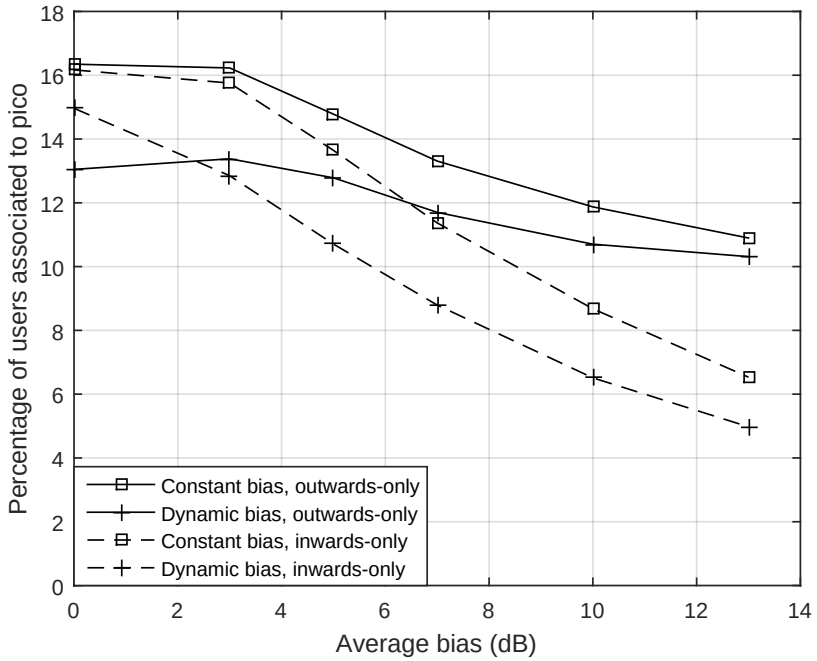

Figure 3.9: Percentage of users associated with the pico with pico QoS.

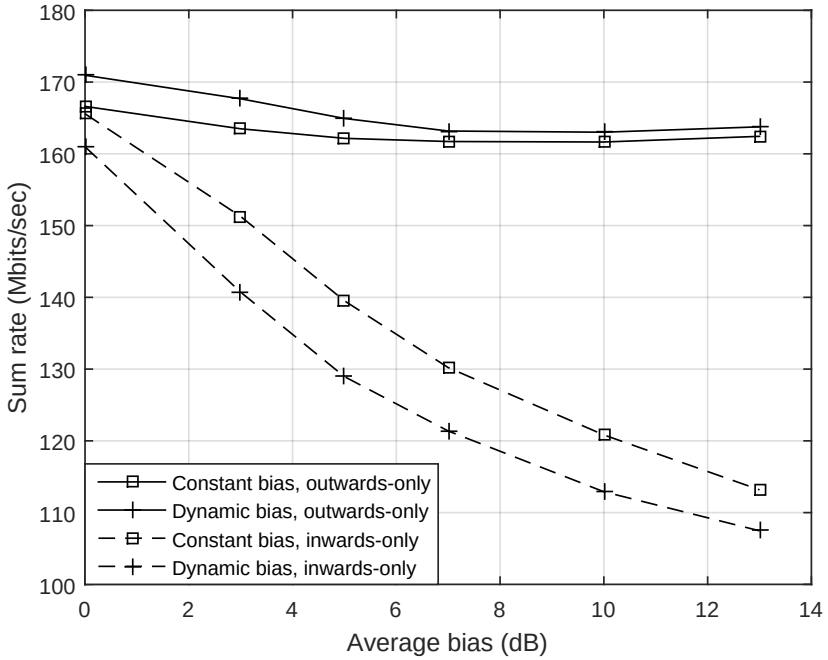

Figure 3.10: Sum rate with pico QoS.

3.4.3 Discussion

The aim of a dynamic bias is to pick the best bias at any given load for a small cell, and therefore we expect a larger bias for low load, and a smaller bias for high load. However, although the tail end of a dynamic bias function is designed to prevent overloading, its benefit is somewhat nullified since even with a constant bias, some users satisfying the biased criteria would have already been associated. For instance, a bias of 4 dB will associate all users with a bias of 3 dB, meaning that if there are no users in between the biased values, then a bias of 4 dB will have the same effect as a bias of 3 dB. A QoS constraint will further limit the benefits of a dynamic bias as it is an artificial overload prevention, hence the converging performances seen in Figs. 3.8 to 3.10.

3.5 Summary

This chapter has studied the concept of load balancing from two perspectives. Firstly, we have investigated in further detail current load balancing schemes, namely biasing and CRE, and analysed the effects of a load dependent dynamic bias function compared to a constant bias value. We have proposed a logistical function with two association orders, outwards-only and inwards-only, and found that under no pico QoS constraints, dynamic biasing with either association order leads to higher average pico rate and total sum rate with the exception of a lower sum rate when inwards-only is used. If pico QoS constraints are in place, the same improvements exist but the performances of dynamic and constant biasing are convergent with larger average bias values. We conclude that dynamic biasing with outwards-only association acts as a suitable natural prevention to pico overloading and can improve overall sum rate.

Chapter 4

Network Balance Index

Key Question: How are **user fairness** and **network balance** different, and what is the benefit of considering network balance instead of fairness?

As networks become more complex, new metrics are needed to quantify performance. The notion of fairness has become increasingly important as it is a better indicator of network management than rate-derived metrics. This is particularly critical for 5G networks, where clustering and heterogeneous user distributions are expected to lead to more non-uniform networks [14, 45].

To ensure quality of service for all users, fairness can be incorporated into user association algorithms [21] by encouraging load balancing via optimising proportional fair utility functions [69]. In this regard, the quantitative measures of fairness used in the literature include Jain's Fairness Index (JFI) and the more generalized α-fairness [31, 69]. In the case of JFI, the fairest network occurs when all users receive the same rate. However, a network where all users receive similar but very low rates (e.g., due to load imbalance and base station being congested) can be deemed 'fair' even though it is undesirable from a load balancing perspective. Thus, in general, fairness metrics cannot be used as a reliable indication of good load balance.

While there have been many efforts in the literature towards methods that facilitate load balancing in HetNets [5], the degree of balance achieved is not explicitly quantified. To the best of the authors' knowledge, there is no formal definition of a balanced network load in the literature, although it has been noted in some papers that *balancing* network load is not necessarily the same as *equalizing* network load [69]. In addition, there is no quantitative measure of network balance, and as such,

no objective or comparative way of determining how balanced a network is (regardless of how the user association is achieved). Addressing these two issues is the main focus of this chapter.

This chapter is organized as follows. We first define a notion of expected load for a network, and consequently propose a network balance index that measures the deviation of an actual load distribution with the expected load. To illustrate the usefulness of the index, a sum rate improvement algorithm is then presented, and its behaviour and trends derived mathematically. Simulation results verify the advantages of our index over conventional fairness. Finally, our main findings are summarized.

4.1 System Model and Problem Formulation

Consider a region with M fixed location base stations and N users, with N_j denoting the number of users associated to base station j. Each base station has transmit power P_j with bias β_j (such that the effective power is $P_j\beta_j$), with the macro bias set to 0 dB. We assume that each base station transmits at its maximum power constantly, and that each user associates with only one base station at a time. We consider a downlink HetNet where each user initially associates with the closest base station, subject to cell-biasing. This corresponds to users associated to their closest weighted Voronoi cells, where the weights are the biased transmit powers, i.e., $w_j = P_j\beta_j$.

We consider the signal-to-noise-ratio, and assume that interference can be dealt with through interference coordination and orthogonal resource allocation. We assume the proportional fair scheduling scheme, and that time and frequency resources are allocated in round robin fashion, such that if a user i is connected to base station j, that user's rate is inversely proportional to the number of users also connected to that base station [26], i.e.,

$$\frac{1}{N_j}\log_2\left(1+\frac{P_j\psi_{j,i}d_{j,i}^{-\alpha}|h_{j,i}|^2}{\sigma_n^2}\right)=\frac{r_{j,i}}{N_j}, \tag{4.1}$$

where $|h_{j,i}|^2$ is the Rayleigh channel gain from base station j to user i, $d_{j,i}^{-\alpha}$ is the

pathloss due to distance $d_{j,i}$ with pathloss exponent α, $\psi_{j,i}$ represents lognormal shadowing with mean 0 dB and variance σ_s^2, σ_n^2 is the Gaussian noise variance and $r_{j,i}$ is the rate without load considerations.

Our performance metric is sum rate, which is the total of all the users' rates in bits/s/Hz. We also impose a maximum user rate such that sum rate and fairness values are not skewed by users associated with lightly loaded base stations.

4.2 Proposed Network Balance Index

As mentioned, fairness alone is not a reliable indication of network balance. The reason is twofold: (i) fairness metrics do not inherently consider network balance as they do not take into account user and base station density and geography, and (ii) fairness metrics do not capture the under or over utilization of base stations. *While fairness measures the relative spread or similarity of user rates (user centric), network balance should measure the even distribution of network resources according to network topology (network centric).* A balanced network is desired from a load balancing perspective. For instance, an overloaded base station may not be an option for new entering users, hence forcing those users to have to connect to some less desirable base station. In addition, if we reduce the load on a congested base station, it can better serve its remaining users.

In order to define the proposed network balance index, we need to first define load and expected (i.e., balanced) load. We define load as the percentage of users associated with a base station. In an ideal balanced network, the expected load of a base station should be (i) proportional to the biased transmit power of each base station, and (ii) inversely proportional to the density of surrounding cells, i.e., the more base stations around a particular cell, the less load that cell should have due to competing base stations. Note that defining expected load must be done prior to any knowledge about user distribution or association. Otherwise, it is possible to construct a definition of network balance that will be high for any arbitrary user distribution.

To mathematically model these properties, we use the area of a multiplicatively

weighted Voronoi cell [70] to represent coverage areas and to help define the expected load. If the area of the weighted Voronoi cell of base station i is x_i, its expected load portion expressed as a percentage is

$$e_i = 100 \times \frac{x_i}{\sum_{j=1}^{M} x_j}. \tag{4.2}$$

Unfortunately, x_i has no known closed form expression [70], but can be estimated by geographical data in real applications (distance distributions exist, but cannot determine generic areas [71]). In our simulations we will use Monte Carlo methods to approximate x_i.

Let $\mathbf{e}$ be a $1 \times M$ vector, with elements e_i from (4.2), denoting the expected load distributions a_i. Let $\mathbf{a}$ be a $1 \times M$ vector containing the actual base station load distribution. We propose a network balance index as follows:

Definition 4.2.1. *The network balance index (NBI) is a measure of the deviation of the current load distribution from the expected load distribution, i.e.,*

$$NBI = 1 - \frac{\|\mathbf{e} - \mathbf{a}\|_1}{200}, \tag{4.3}$$

where $\|\cdot\|_1$ is the ℓ_1 norm. The vector $\|\mathbf{e} - \mathbf{a}\|_1$ is divided by 200 because the maximum possible deviation is 200% (deviation will range from maximum -100% for total overloaded or+100% for total underloaded). Like JFI, NBI has values in the range $[0, 1]$.

To illustrate, consider a network with one macro and three picos, and that their expected load distributions are $[40, 20, 20, 20]$. If the actual distributions are $[80, 10, 10, 0]$, then the NBI would be

$$1 - \frac{|-40| + |10| + |10| + |20|}{200} = 1 - \frac{80}{200} = 0.6, \tag{4.4}$$

indicating that the actual network is 60% balanced with respect to the expected load distribution.

Network balance can provide information about the network that fairness alone does not. For instance, even if sum rate has been optimized with a minimum fairness,

improving this sum rate further cannot be done without knowing whether to increase or decrease the fairness constraint. With network balance however, we can identify certain scenarios where increasing NBI will also lead to increasing fairness.

4.3 Sum Rate Improvement Algorithm Using NBI

To show an application of the proposed metric in network planning, we propose an algorithm that uses NBI as an indicator to refine the initial user association. Note that a refinement algorithm cannot aim to increase sum rate, as sum rate is not normalized and therefore optimal values cannot be known beforehand. Thus, we aim to increase NBI, as it can identify which base stations are overloaded and which ones should receive offloaded users.

Our algorithm first associates users to the base station of the weighted Voronoi cell they are located in and then computes the NBI. Next, we denote the most overloaded base station and its users by $s_i \in \mathcal{O}$, $i = 1, \cdots, |\mathcal{O}|$, and the closest empty base stations followed by the most underloaded base station (ordered from most empty/underloaded to least if there are multiple), by $\mathcal{U}_j$. At each step, the user from $\mathcal{O}$ that is closest to $\mathcal{U}_j$ is re-associated, such that

$$\left\{ s_i \notin \mathcal{O}, s_i \in \mathcal{U}_j | a_j \leq e_j \right\}. \tag{4.5}$$

Each base station $\mathcal{U}_j$ will gain a user from $\mathcal{O}$ until $\lfloor \theta N \rfloor$ users are re-associated, or until any $\mathcal{U}_j$ reaches its expected load e_j. $\theta \in [0, 1]$ is the maximum fraction of total users that can be offloaded, rounded down to the nearest integer (line 4 in Algorithm table).

4.3.1 Condition for Increasing NBI and Sum Rate:

We can analytically show for a clustered user network that our algorithm increases sum rate by increasing NBI through offloading users. Let $r_{\{\text{over,under}\},i}$ be the rate received by the ith user associated with an overloaded or underloaded base station without load considerations. Suppose one user is being offloaded from an overloaded

Algorithm 4.1 Re-associating users based on NBI.

1: Associate users using (biased) min-distance association.
2: Determine most overloaded ($\mathcal{O}$) and most empty and underloaded base stations ($\mathcal{U}_j$).
3: **Initialize** re-associate $= 0$
4: **while** re-associate $< \lfloor \theta N \rfloor$ and $a_j < e_j$ **do**
5: Determine user $s_i \in \mathcal{O}$ closest to $\mathcal{U}_j$.
6: Offload user from $\mathcal{O}$ to $\mathcal{U}_j$, i.e., $s_i \notin \mathcal{O}, s_i \in \mathcal{U}_j$.
7: re-associate = re-associate + 1.
8: Move onto next $\mathcal{U}_j$
9: **end while**

base station with N_{over} initial users to an underloaded base station with N_{under} initial users. Using (4.1), for sum rate to improve, the difference in sum rate before and after offloading must be greater than 0, i.e.,

$$\underbrace{\left(\sum_{i=1}^{N_{\text{over}}-1}\frac{r_{\text{over},i}}{N_{\text{over}}-1}+\sum_{i=1}^{N_{\text{under}}+1}\frac{r_{\text{under},i}}{N_{\text{under}}+1}\right)}_{\text{After offloading}}-\underbrace{\left(\sum_{i=1}^{N_{\text{over}}}\frac{r_{\text{over},i}}{N_{\text{over}}}+\sum_{i=1}^{N_{\text{under}}}\frac{r_{\text{under},i}}{N_{\text{under}}}\right)}_{\text{Before offloading}}>0. \quad (4.6)$$

It is clear that the rates of those $N_{\text{over}}-1$ users still associated with the overloaded base station will improve, since

$$\sum_{i=1}^{N_{\text{over}}-1}\frac{r_{\text{over},i}}{N_{\text{over}}-1}>\sum_{i=1}^{N_{\text{over}}-1}\frac{r_{\text{over},i}}{N_{\text{over}}}. \quad (4.7)$$

Therefore, for sum rate to improve, we require that

$$\underbrace{\sum_{i=1}^{N_{\text{under}}+1}\frac{r_{\text{under},i}}{N_{\text{under}}+1}-\sum_{i=1}^{N_{\text{under}}}\frac{r_{\text{under},i}}{N_{\text{under}}}}_{\triangleq\Omega}-\frac{r_{\text{over},N_{\text{over}}}}{N_{\text{over}}}>0. \quad (4.8)$$

If $N_{\text{under}} \to \infty$, then $\Omega \to 0$, leading to (4.8) becoming false. Therefore, we conclude that (4.8) is most easily satisfied if N_{under} is small and N_{over} is large, which is exactly the case when users are heavily clustered. Note that for the special case when the underloaded base station is initially empty, i.e., $N_{\text{under}} = 0$, (4.8) reduces to

$$r_{\text{under},1}>\frac{r_{\text{over},N_{\text{over}}}}{N_{\text{over}}}, \quad (4.9)$$

which is more easily satisfied when $N_{\text{over}} \to \infty$.

4.3.2 Relationship between Sum Rate and Fairness:

We can obtain mathematical insight into how an increase in sum rate may affect JFI. Using the JFI definition [31], the difference in fairness before and after a user is offloaded is

$$\frac{\left(A + \sum_{i=1}^{N_{\text{over}}-1} \frac{r_{\text{over},i}}{N_{\text{over}}-1} + \sum_{i=1}^{N_{\text{under}}+1} \frac{r_{\text{under},i}}{N_{\text{under}}+1}\right)^2}{N\left(B + \sum_{i=1}^{N_{\text{over}}-1}(\frac{r_{\text{over},i}}{N_{\text{over}}-1})^2 + \sum_{i=1}^{N_{\text{under}}+1}(\frac{r_{\text{under},i}}{N_{\text{under}}+1})^2\right)} - \frac{\left(A + \sum_{i=1}^{N_{\text{over}}} \frac{r_{\text{over},i}}{N_{\text{over}}} + \sum_{i=1}^{N_{\text{under}}} \frac{r_{\text{under},i}}{N_{\text{under}}}\right)^2}{N\left(B + \sum_{i=1}^{N_{\text{over}}}(\frac{r_{\text{over},i}}{N_{\text{over}}})^2 + \sum_{i=1}^{N_{\text{under}}}(\frac{r_{\text{under},i}}{N_{\text{under}}})^2\right)}, \tag{4.10}$$

where A is the sum of rates for the other users, and B is the sum of their squared rates.

If users are very clustered, i.e., N_{under} is small and N_{over} is large, we have already established that sum rate tends to increase. Further, that increase is almost solely attributed to the offloaded user, since if N_{over} is large we assume that the $N_{\text{over}} - 1$ users of the offloading base station experience similar rates as before. Thus, we can approximate (4.10) as

$$\frac{(A' + \epsilon)^2}{N\,(B' + \epsilon^2)} - \frac{(A')^2}{N\,(B')}, \tag{4.11}$$

where A' and B' are new constants and ϵ is the increase in sum rate. If we set (4.11) to 0 to study under which conditions it will be positive or negative, the numerator can be reduced to

$$2A'B' + \epsilon(B' - (A')^2) = 0. \tag{4.12}$$

Since $(A')^2$ is a square of sums, and B' is a sum of squares, $(B' - (A')^2) \leq 0$. Therefore, when ϵ is large, which tends to be the case when users are initially highly clustered, fairness tends to decrease since (4.12) will become a '$<$' inequality. Conversely, as ϵ becomes smaller, fairness tends to increase.

4.4 Simulation Results

We simulate a network with one macro in the centre of a 1 km $\times$ 1 km region, with 4 small cells spaced at a radius 250 m around the macro. Each small cell could be a pico or a femto and this is randomly generated in each realization. Bias values are 7 dB and 13 dB for picos and femtos respectively, while transmit powers are 40 dBm, 30 dBm and 20 dBm in order of decreasing tiers. Pathloss exponent is $\alpha = 3$, lognormal shadowing variance σ_s^2 is 6 dB, noise power σ_n^2 is -174 dBm/Hz, and the transmission bandwidth is 20 MHz. Users are randomly distributed according to a Thomas cluster process[2] centered at the macro with intensity function [72]

$$\lambda(\mathbf{x}) = \frac{\bar{c}}{2\pi\sigma^2} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right), \tag{4.13}$$

where the average number of users is $\bar{c} = 500$, σ^2 is the cluster variance, $\mathbf{x}$ is the position vector of the user relative to the parent point, and $\|\cdot\|$ denotes Euclidean norm. The maximum rate for each user is limited to 2 bits/s/Hz, and we set $\theta = 0.1$. To calculate the Voronoi cell area x_i, we divide the region into a 100×100 grid and determine how many grid elements each Voronoi cell contains.

4.4.1 Sum Rate Improvement

We first compare the sum rate performance of our proposed algorithm with minimum distance association (where users associate with the closest base station) and the dynamic range heuristic proposed in [26], which aims to select the best number of users to connect to picos. Since the system model in [26] is different, for fair comparison we have adapted the heuristic to our system model, such that each user compares the received powers from its nearest pico and the macro, and at most $\lfloor \theta N \rfloor$ users can be re-associated.

From Fig. 4.1 we observe that our proposed algorithm drastically improves the

[2]Our NBI definition is independent of the user distribution. Thomas cluster process is adopted as an example of a clustered process. Other clustered processes display similar performance. Uniformly distributed processes such Point Poisson Process display minor improvements in sum rate and NBI, and thus their results are not shown.

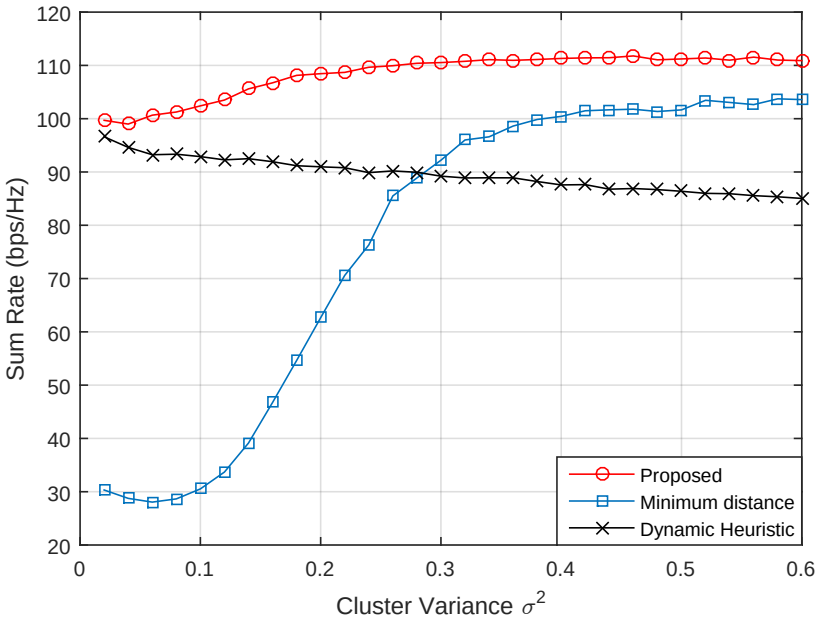

Figure 4.1: Sum rate of three user association schemes - minimum distance, dynamic heuristic and proposed algorithm with varying Thomas cluster variance.

sum rate compared to conventional minimum distance association, and slightly outperforms the dynamic range heuristic for all cluster variances. Interestingly, while [26] shows that its dynamic range heuristic is very close to optimal association, this claim is only valid for their system model where a user as the option of connecting to one macro or one pico. Since our system model contains multiple small cells for the user to choose from, our proposed algorithm outperforms the dynamic range heuristic as it takes load balancing into account by deciding which small cells should receive which re-associated user.

4.4.2 Average Improvement

Fig. 4.2 shows the average improvement of our proposed algorithm over conventional minimum distance association and dynamic range heuristic after 100 realizations for varying σ^2. User locations, channel fading and small cell types are all varied for each realization. Our algorithm always improves NBI, but when users are heavily clustered (small σ^2), sum rate is drastically improved despite decreasing JFI. As users become more uniform, sum rate improvement decreases, at which point both NBI

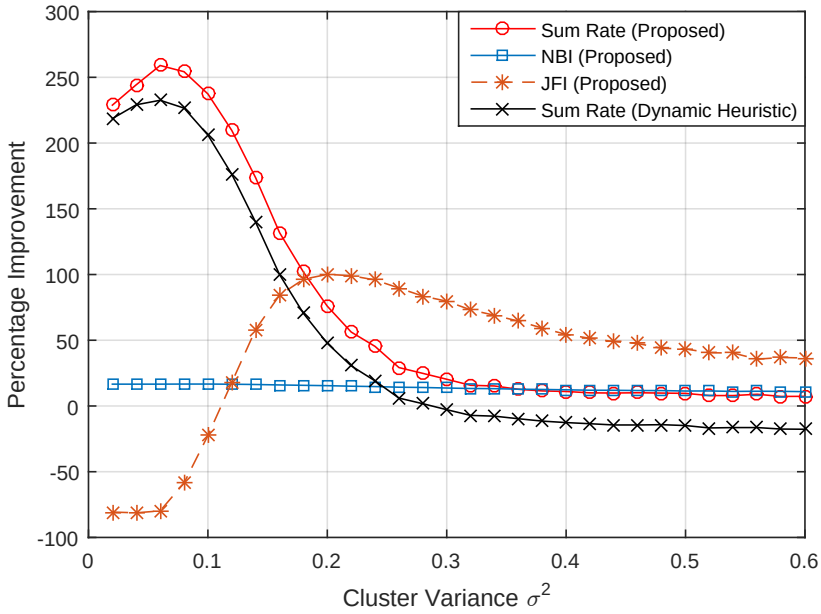

Figure 4.2: Percentage improvement in sum rate using proposed algorithm compared to conventional minimum distance association and dynamic heuristic with increasing Thomas cluster variance. Percentage improvements in NBI and JFI with proposed algorithm are also shown.

and JFI will increase, and continue to do so even as users become more uniformly located. This is due to the fact that larger initial clustering leads to offloaded users gaining more from associating with a less loaded base station. From another perspective, higher clustering means more overloading and greater potential improvement, while more uniformity means less overloading and less improvement.

For more uniform user distribution (larger σ^2), offloaded users will benefit much less, if at all, hence maintaining sum rate. The observations from our analysis are also verified by these results, as the largest increases in sum rate occur when clustering is high (small σ^2) as suggested by (4.8), which also approximately coincides with the largest decreases in fairness. *This behaviour suggests that using the NBI as an indicator is most beneficial when users are heavily clustered.* As users become more uniformly located, fairness begins to increase from our algorithm, as higher rates experienced by users associated with the underloaded cell are brought down closer to those of clustered users. This behaviour is consistent with that described by (4.12), as a reduced sum rate improvement (i.e., smaller ϵ) leads to increasing fairness.

4.5 Summary

In this chapter, we have provided a new perspective about load balancing by proposing a new metric, the network balance index, that quantifies the amount of balance in cellular HetNets. We have described how network balance is conceptually different to user fairness, and have shown how a user association algorithm can use this as a means to improve sum rate and fairness. Our NBI can be exploited to increase sum rate despite decreasing fairness when users are heavily clustered, and maintain sum rate while increasing fairness when users are more uniform. Future work can explore directly incorporating NBI in optimizing user association.

Chapter 5

Preference Association and Network Dynamics

Key Question: *Can network states be predicted if there are minor changes? Which users are* ***mostly likely to change associations****?*

User association is a critical process in heterogeneous networks (HetNets) and cellular communications that connects users to suitable base stations as a means of accessing the network [21, 66]. In the literature, user association has been studied using many mathematical approaches, including as a matching problem with parallels to the college admissions game in game theory [27, 73], as an optimization problem [23, 24], or modeled as a stochastic game [74]. Future networks with more dense user and base station deployment will require a deeper understanding of user association mechanisms and more complex schemes to provide the best quality of service to each user. In this regard, conventional user association has aimed to improve a system utility, most commonly sum rate or capacity. Fairness is a popular alternative metric to consider, though often as a secondary thought, e.g., imposed as a constraint rather than an objective to maximize. However, high fairness (in terms of a quantitative measure such as Jain's Fairness Index (JFI)) may be beneficial in scenarios where users might be accessing the same information from the network, and therefore may become a primary objective in user association for future networks.

In addition to achieving desirable quantitative metrics such as fairness and sum rate, a study of the effects of network dynamics on user association and network state is particularly important for future networks. For instance, if the number of users in a large network changes slightly, e.g., one user enters or leaves, the new

association may be very similar to the previous, and therefore it is unnecessary to recalculate all associations. Predicting or at least determining probabilities of future user associations to decrease computation can be beneficial to large networks. In this regard, sequential or predictive user association was mentioned in [5], and a similar concept was briefly considered in [75] but with a specific and rigid system model. Although game theory seems to be a suitable technique to study the interaction between user and network behaviour and can lead to suitable strategies to achieve an objective [28], current research generally models each instance (i.e., fixed number of players) as a game or Markov decision process, rather than what happens when the number of players change [76]. Further, game theory is not an analytical tool and thus is limited in its ability to explain trends or predict behaviours. A user entering or leaving a base station is studied in [29] and the Nash equilibrium property is proven to hold, though the focus is on user rates.

The number of base stations in the network may also change depending on factors such as energy saving and network load. Base stations can be turned off during off peak times to save power [77], while ad hoc base stations such as ones deployed through drones may be used to serve hotspots or bursty traffic [78]. Entering or exiting base stations can dramatically affect the network state, and consequently a study of network behaviour with base station dynamics can bring insight into network design. To the best of our knowledge, the study of user association to guarantee high fairness in dense networks, including network dynamics such as users and/or base stations entering and exiting the network, is an important open problem.

This chapter can be organized as follows. We first propose a base station preference association scheme, and mathematically prove that generally this leads to high rate fairness among users assuming round robin resource scheduling. Next, we derive the associated rank distribution and probabilities of re-association when a single user or base station enters or leaves the network, and describe the effect of key parameters on these probabilities. Simulation results verify that there exists a type of user that are most likely to re-associate with network dynamics. Finally, we summarize our main findings.

5.1 System Model and Preference Association

Consider a HetNet with M randomly located base stations, which could be a mixture of macro and small cells. N users are also randomly distributed.

Suppose each user calculates the average receive powers from each base station, given by

$$P_j d_{i,j}^{-\alpha}, \tag{5.1}$$

where P_j is the transmit power of the jth base station, $d_{i,j}$ is the distance in metres between the ith user and jth base station, and α is the pathloss exponent.

Each user then feeds back the received powers to the corresponding base stations, and the base stations then construct their preference lists, i.e., ranking each user according to the received powers. Each list represents the users each base station would most like to associate with.

5.1.1 Base Station Preference Association

We define the base station preference association rule as users associating with the base station where it is ranked highest in. The value of this highest rank is termed the *associated rank* of that user. The number of users that are higher than a particular user at its associated rank is $0 \leq K_i \leq N-1$, such that its associated rank is $K_i + 1$.

Note that for a HetNet with small cells and different base station transmit powers, this association rule will be different to max received power association, as it is possible for a user to have a smaller received power from a femto than from a macro, but be ranked higher in the femto's preference list. If a user is ranked equally highest in multiple base station lists, the user will randomly pick any of those *tied* base stations to associate with. The number of ties for user i is $0 \leq A_i \leq M$. Fig. 5.1 illustrates the defined parameters. The circled users are the same user appearing in each base station's list.

To define various terminology, we describe a *weakly associated* user as one that has multiple ties for its highest rank (with *more weakly associated* meaning a larger A_i). A *strongly served* user is one where its highest rank is high up in the preference

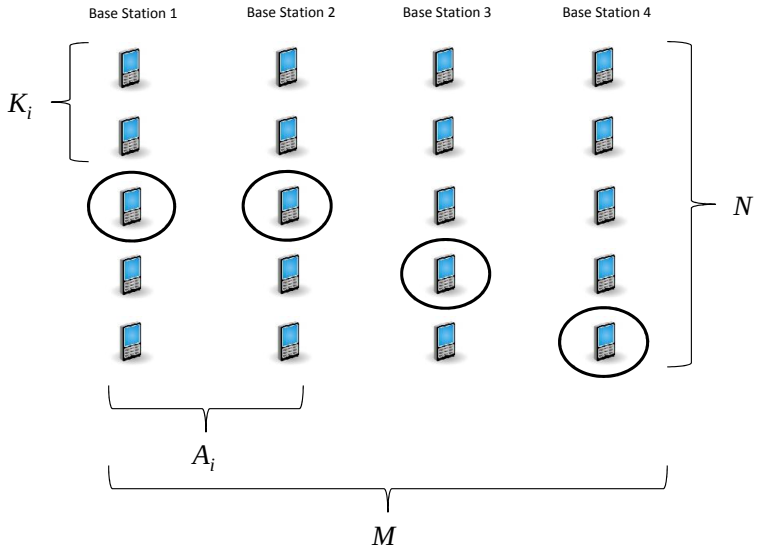

Figure 5.1: Definitions of K_i, A_i, M and N.

list (small K_i), while a *weakly served* user is one where its highest rank is low in the preference list (large K_i).

This preference association was inspired from [73], with the analogy of colleges to base stations and students to users. However, while in the college admissions both colleges and student's preferences are considered, in this association rule, associations are made from the base station's perspective, and does not consider each user's own base station preference. This subtle difference is important as our aim is to improve overall network performance, rather than to ensure the maximization of each independent user's intentions.

5.2 Fairness Analysis

The most intuitive notion of fairness is one where maximum fairness is achieved when all users achieve equal rates, and minimum fairness when all users experience different rates. Though this basic idea of fairness may not be suitable for all scenarios, our study makes no assumptions regarding the relative requirements of base stations or users, thus necessitating the use of this fairness notion. The accepted metric to

quantify this fairness is the JFI [31], which is defined as

$$\frac{\sum_{i=1}^{M}(r_i)^2}{M\sum_{i=1}^{M} r_i^2}, \tag{5.2}$$

where

$$r_i = \log_2\left(1 + \frac{P_j d_{i,j}^{-\alpha}|h_{i,j}|^2}{\sigma^2}\right) \tag{5.3}$$

is the rate for the ith user without load considerations, $h_{i,j}$ is the Rayleigh channel coefficient and σ^2 is the additive white Gaussian noise power.

Due to the $\log_2$ term, even if the transmit powers are orders of magnitude apart, the actual rate values would be within the same order of magnitude. Therefore, fairness of rates for each user without load considerations would be high, i.e., close to 1.

However, actual rates users experience are dependent on the load of their associated base station, meaning that rates are divided by the number of users also associated with that base station, assuming round robin scheduling of time and frequency resources, i.e.,

$$r_i = \frac{1}{N_j}\log_2\left(1 + \frac{P_j d_{i,j}^{-\alpha}|h_{i,j}|^2}{\sigma^2}\right), \tag{5.4}$$

where N_j is the number of users associated to base station j.

It is easy to see that conventional association rules such as maximum received power or minimum distance may result in unbalanced load distributions. Since the load is a pre-log term, large differences in load will drastically *reduce* rate fairness. In other words, to *maintain* high fairness, it is desirable to have base stations with equal load, i.e, similar number of users associated to them. Mathematically, this is due to the scale invariance property of the JFI - if all rates are scaled by the same factor, JFI will remain the same.

5.2.1 Proof of High Fairness for Preference Association

If either base stations or users are randomly distributed, then the probability that a particular user appears in a specific rank on a base station's preference list is $\frac{1}{N}$.

For a particular base station, the probability that a user at position i will be ranked lower in all other base station lists (i.e., the user is associated to that base station) is

$$\left(\frac{N-i}{N}\right)^{M-1}, \tag{5.5}$$

since there are $N-i$ positions lower than position i, and $M-1$ other base station lists where this must be true.

Since there are N users, and each user has a $\frac{1}{N}$ chance of being in any of i positions in a base station's list, the expected number of users associated to each base station is therefore

$$\begin{aligned}\frac{1}{N}N\sum_{i=1}^{N}\left(\frac{N-i}{N}\right)^{M-1} &= \frac{1}{N^{M-1}}\sum_{i=1}^{N-1} i^{m-1} \\ &\overset{\text{(a)}}{=} \frac{1}{N^{M-1}}\left(\frac{N^M}{M}-\frac{N^{M-1}}{2}+\frac{(M-1)N^{M-2}}{12}+\mathcal{O}(N^{M-4})\right) \\ &= \frac{N}{M}-\frac{1}{2}+\frac{M-1}{12N}+\mathcal{O}(N^{-3}),\end{aligned} \tag{5.6}$$

where $\mathcal{O}(N^{-3})$ is a polynomial in N of at most degree -3. The equality (a) is obtained from [79].

For realistic N users and M base stations ($N > M$), $\frac{N}{M}$ will be the dominant term. Thus, users are divided approximately evenly across all base stations, leading to even load distribution and high rate fairness.

5.2.2 Distribution of Associated Ranks

The distribution of associated ranks is not uniform, as more users will have a lower-valued associated rank than a higher-valued one.

For a user at rank K_i+1, there is a $\left(\frac{N-K_i}{N}\right)^{M-1}$ probability that its rank in the other $M-1$ base stations will be equal or lower to it. Thus, with M base stations,

there are

$$M\left(\frac{N-K_i}{N}\right)^{M-1} \tag{5.7}$$

users who should have $K_i + 1$ as their associated rank.

5.3 Association Probabilities with Entering or Exiting Users

If we consider how the associations will change with entering or exiting users, we note that with a single entering or exiting user, the rankings of an existing user will change by at most one position. Therefore, we can deduce that only users that are weakly associated, i.e., have ties where their highest ranking belongs to multiple base stations, will have to switch associations.

A current user's ranking will only change if an entering or exiting user is ranked above, since an entering or exiting user ranked below would not change the ranks of any users above and hence have no effect on the association decision.

5.3.1 Entering User

If an entering user is ranked above user i in its associated base station's list, user i will be pushed down from rank $K_i + 1$ to $K_i + 2$, meaning that it may re-associate to one of its tied base stations, provided that in *none* of them is the entering user also ranked above. Thus, the probability that user i will have to change its association due to an entering user is

$$X_{enter} = \frac{K_i+1}{N-1}\left(1-\left(\frac{K_i+1}{N-1}\right)^{A_i-1}\right). \tag{5.8}$$

The first term $\frac{K_i+1}{N-1}$ is the probability that user i is pushed down in its current associated base station list ($N+1$ because there are that many positions in the list a new user can enter into), while the second is the probability that at least one tied rank out of $A_i - 1$ base station lists retains its position.

For fixed and given N, M and A_i, we can find the most likely position where a

user may have to re-associate by differentiating (5.8) with respect to K_i:

$$\frac{\mathrm{d}X_{enter}}{\mathrm{d}K_i} = \frac{1}{N+1} - \frac{A_i(K_i+1)^{A-1}}{(N-1)^{A_i}}. \tag{5.9}$$

Setting the above to 0,

$$K_i = \sqrt[A-1]{\frac{(N+1)^{A_i-1}}{A_i}} - 1. \tag{5.10}$$

Thus, the most likely position for a user to re-associate is $K_i + 1$, rounded to the nearest integer.

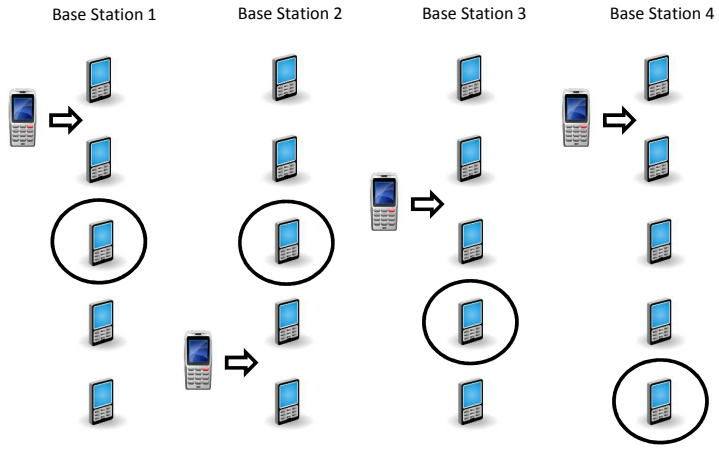

Figure 5.2: User entering network. If circled user was initially associated with base station 1, it will now associate with base station 2 since the user in base station 1's list has been pushed down.

5.3.2 Exiting User

A current user i will only need to re-associate to another tied base station if an exiting user ($N-1$ possible users) from its associated base station list exists from the last $N - K_i - 1$ ranks (user i maintains its rank), and at least one other tied base station has the exiting user exit from their first K_i ranks (such that at least one tied base station rank gets pushed up). The probability of this occurring, and hence forcing user i to re-associate, is

$$X_{exit} = \frac{N-K_i-1}{N-1}\left(1 - \left(\frac{N-K_i-1}{N-1}\right)^{A_i-1}\right). \tag{5.11}$$

The first term $\frac{N-K_i-1}{N-1}$ is the probability that user i retains its current associated base station rank, while the second is the probability that at least one tied position out of $A_i - 1$ tied base station lists is pushed up.

Verifying when $K_i = 0$, $X_{exit} = 0$, meaning that user will never have to change associations.

The most likely positioned users to re-associate when a user exists can be determined by:

$$\frac{\mathrm{d}X_{exit}}{\mathrm{d}K_i} = \frac{A_i(N-K_i-1)^{A-1}}{(N-1)^{A_i}} - \frac{1}{N-1}. \tag{5.12}$$

Setting the above to 0,

$$K_i = N - 1 - \sqrt[A-1]{\frac{(N-1)^{A_i-1}}{A_i}}. \tag{5.13}$$

Generally, (5.10) is a smaller value than (5.13).

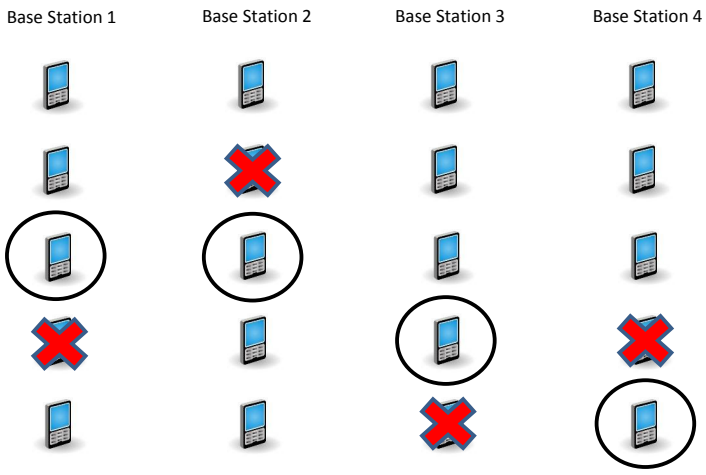

Figure 5.3: User exiting network. If circled user was initially associated with base station 1, it will now associate with base station 2 since the user in base station 2's list has been pushed up.

5.3.3 Effect of A_i, K_i, and N on Association Probability

Plotting each of the parameters while keeping the others fixed shows both expected and unexpected trends.

5.3.3.1 Varying A_i

Increasing the number of ties for a particular user i, i.e., more weakly associated, increases the probabilities of changing association in (5.8) and (5.11). This is expected as weakly associated users have more options to re-associate, and are more likely to do so if there is a change in the network.

5.3.3.2 Varying K_i

Interestingly, there exists a value of K_i where a user of that position is more likely to re-associate than any other. This position is neither a strongly nor a weakly served user, but an average served user. The value of this K_i is different for (5.8) and (5.11) as shown in the preceding section.

This observation can be explained as follows. Strongly served users will not likely need to re-associate, as they have little to benefit from a single entering or exiting user and load balancing. Weakly served users would also not likely to re-associate as they are likely to receive the same perceived benefits from any re-association, meaning that they are still more likely to have their highest ranking with their current base station.

We note that this finding makes intuitive sense, as it parallels with real life college admissions. Strong candidates will likely to preferred by most or all colleges, and therefore have little incentive to change their own preferences, as will weak candidates who will stay unpreferred even if some new candidates appear. However, average candidates, who might be preferred by some colleges but not by others, are most volatile in terms of their decision, and will be most affected by any change in the process.

5.3.3.3 Varying N

Increasing the number of users in the whole network decreases the probabilities of changing association in (5.8) and (5.11). This is expected since larger networks appear more similar to each other than smaller networks, meaning that associations are more likely to remain unchanged.

5.4 Association Probabilities with Entering or Exiting Base Stations

Because the addition or removal of a base station does not affect the ordering of existing preference lists, base station dynamics is simpler to analyse than user dynamics.

5.4.1 Entering Base Station

With the inclusion of an additional base station, the chance of any particular user switching association to the new base station is $\frac{K_i}{N}$, since there are K_i positions in the new base station's preference list that the i user can be ranked in and switch association to. Therefore, the total number of expected users that may re-associate to the new base station is

$$\sum_{i=1}^{M} \frac{K_i}{N}. \tag{5.14}$$

5.4.2 Exiting Base Station

If the number of base station reduces by one, all users associated with the exiting base station would simply associate with the base station where it was ranked second highest in, or with one of the $A_i - 1$ tied base stations.

5.5 Simulation Results

In our simulations we randomly distribute $M = 5$ base stations with $N = 100$ Point Poisson Process users. The base stations are randomly chosen to be either a macro, pico or femto, and all transmit at their maximum powers of 40, 30 and 20 dBm respectively. We set the pathloss exponent to $\alpha = 2$. Channels are Rayleigh fading with mean 0 and noise power $\sigma^2 = -174$ dBm/Hz. Fig. 5.4 plots the simulated and theoretical distribution of associated ranks. The two plots match almost perfectly, confirming that associated ranks are concentrated towards lower rank values.

Fig. 5.5 compares the JFI of all user rates from different user association schemes, including maximum received power, nearest base station, and a dynamic range heuristic from [26]. As discussed in Section 5.2, rates without load consideration

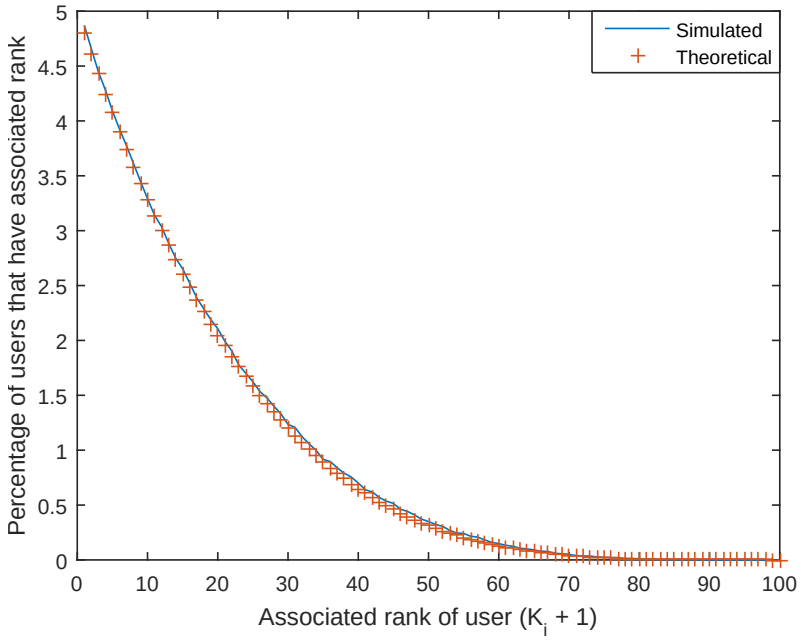

Figure 5.4: Distribution of associated ranks. Associated ranks are not uniformly distributed, but are concentrated towards smaller values.

naturally have a high fairness value. Once load is taken into consideration, our proposed preference association maintains a high JFI, while other associations significantly decrease fairness. If fairness is an important factor, preference association is a much more suitable method compared to conventional associations.

Fig. 5.6 plots the percentage of times a user of a particular associated rank will change associations with a single entering or exiting user. Both scenarios show that there exists a rank where users are most likely to re-associate. The results shown here are generated with no fixed A_i, and therefore do not correspond exactly to (5.8) and (5.11). However, we observe that the peak of exiting user probability occurs to the right of (i.e., higher valued associated ranks) entering user probability. This can be explained by jointly considering Sections 5.2.2 and 5.3.

Our analysis shows that the most likely K_i is generally larger for exiting users than entering users, and hence one would expect a negative skew shape for exiting users, and a positive skew shape for entering users. However, Fig. 5.6 shows a

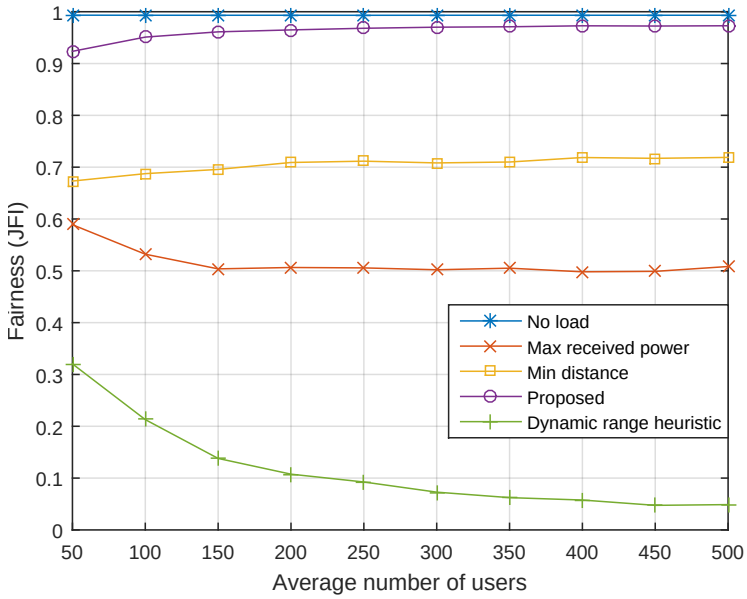

Figure 5.5: Fairness of user rates for various association rules with random base station locations and PPP users.

positive skew for both, which is a result of multiplying individual associated ranks by (5.7). Since there are more users with a smaller valued associated rank, the final shape of exiting user probabilities will be weighted towards the lower valued ranks, and therefore still be positively skewed. The sum of these probabilities (i.e., area under the plot) multiplied by the total number of users gives the expected number of re-associations due to a single entering or exiting user.

In addition, Fig. 5.6 shows that exiting users have more of an effect on the network than entering users, which agrees with our intuitions. Larger networks look and behave more similar to each other than smaller networks. Also, as indicated by (5.11), no re-associations can occur for users with associated ranks of 1.

5.6 Summary

We have proposed a new base station preference association scheme where users associate with the base stations who prefer it the most. Our analysis and simulation

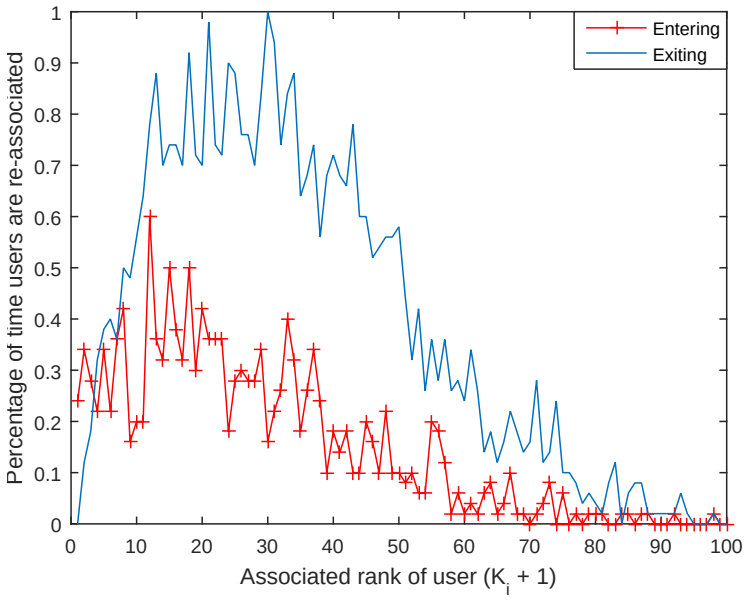

Figure 5.6: Percentage of times user of a particular associated rank re-associated due to a single entering or exiting user. Exiting users induces more change in user association than entering users.

results confirm that it leads to high fairness compared to conventional association schemes. Using this preference association, we have studied the effects of network dynamics, namely entering and exiting single users and base stations, on the user association state. Re-association probabilities and associated rank distributions are derived and verified by simulations. Our results indicate that a typical weakly associated user is mostly likely to re-associate given a network dynamic, and that shrinking network size has more effect on user association than a growing network.

Chapter 6

D2D Mode Selection and Resource Allocation

Key Question: *How might a network **decide** to allow D2D, and if it does, which **mode** and **parameters** should it choose?*

D2D communications allowing direct communication between nearby users has been envisaged in 3GPP standards [80]. From an operator perspective, determining the type of D2D operation during *mode selection* (assuming that neighbor discovery has already been achieved [37]) is a crucial initial decision by the network and an important research topic. In dedicated or cellular mode, the fundamental research challenge is resource allocation. In the reuse mode, the fundamental research challenge is interference management via efficient power control. Overall, in order to provide operator managed quality of service guarantees, centralized solutions which have low complexity are desirable.

Mode selection schemes have been proposed in the literature based on minimum distance between the D2D transmitter (DTx) and D2D receiver (DRx) [81], biased D2D link quality and whether it is at least as good as the cellular uplink quality [82] or guard zones protecting the MBS [83] or D2D users [84]. A limitation of the schemes in [81, 82, 84] is that they do not inherently protect the D2D link from interference, while the scheme in [83] does not impose any restrictions on the D2D distance; generally D2D communication is envisaged as short range direct communication. Also in [84], the guard zone region surrounding D2D users is primarily used to determine which cellular users are allowed to reuse resources allocated to the D2D users, rather than specifically a mode selection criterion. Mode switching and mixed mode

approaches, where multiple modes are utilized at once, are also studied in [85, 86].

Once a mode is decided, the network must address *resource allocation* to meet network requirements. In the reuse mode, *power control* is used to manage transmit powers and hence interference. Power control is not guaranteed to provide closed form analytical solutions, but it has been shown that optimal solutions can at least be found from searching from a finite set [39], although this claim has only been made with two transmitting sources in the system model. In [87], power optimization for one D2D transmitter and one cellular user transmitting during uplink was studied. Since there are two transmitters, the optimization is a simple two-dimensional problem. Power allocation for maximizing sum rate was also studied in [88], where the authors focused on a binary power decision, i.e., powers operate either at their maximums or minimums, and with no SINR guarantee for any user. The authors showed that binary power control is optimal for two users, but is suboptimal for arbitrary number of users.

Orthogonal resource allocation for D2D was studied in [39] for both dedicated and cellular modes using the downlink (DL). In each mode, time and frequency allocation was considered, and for each allocation, greedy (unconstrained) and rate constrained optimization was presented. However, in the rate constrained case, only the cellular user has a minimum rate requirement, and thus the possibility exists for the cellular user to be allocated all the resources and leaving the D2D with none. Further, [39] only considered a single-tier network in its system model. In a two-tier cellular network, a licensed femtocell changes the way resources can be allocated, and in turn changes the maximization of the optimization objective in a non-trivial manner. Note that some papers use joint optimization [40, 89, 90] and/or game theory [91, 92, 93, 94, 95] to solve resource allocation problems. In theory, joint optimization solutions could be optimal, but their complexity often means approximations are required in practice. Further, in two-tier networks, different users may have different constraints or requirements which will further increase the difficulty of finding optimal solutions. Meanwhile, game theory has the advantage that it is a more distributive approach, but does not provide operator managed quality of service guarantees.

Both uplink (UL) and downlink (DL) spectrum resources can be used by in-band D2D. In the literature, there exists works which either use UL [40, 87, 96] or DL [97, 98], and also some which consider both [39, 99]. Generally, interference scenarios are less severe in the UL [34, 92, 100]. However, in this paper we assume DL resources are reused as this represents the worst case interference scenarios.[3]

In summary, existing works on D2D communications have generally considered mode selection, resource allocation and power control sub-problems either separately or considered a subset of these problems for single and multi-tier cellular networks [39, 40, 81, 82, 83, 84, 87, 88, 89, 91, 92, 96, 97, 98, 99]. To the best of our knowledge, a centralized solution for mode selection, resource allocation and power control in D2D-enabled two-tier cellular networks is still an open problem.

The chapter is organized as follows. We first describe a mode selection framework that uses both an interference and distance criteria to decide a suitable D2D mode for a potential D2D pair. Next, for each mode, we show how to determine optimal or near optimal parameters to maximize system sum rate under rate constrained or unconstrained cases. We use simulation results to verify the benefits of our proposed methods, and discuss the scalability of our approach. Finally, we summarize our main findings.

6.1 System Model

We consider a single cell in a two-tier cellular network, as illustrated in Fig. 6.1. Our system model is comprised of: (i) an MBS located at the center of the cell, which is serving a single cellular user equipment (CUE), (ii) an FAP serving a single femto user equipment (FUE), and (ii) a D2D pair comprising of a DTx and a DRx located close to each other. All the different user equipments (UEs), MBS and FAP are equipped with single omni-directional antennas. We assume that suitable inter-cell interference control mechanisms, such as fractional frequency reuse, are employed to avoid or manage inter-cell interference [101]. Hence, we study the single cell

[3]The methodologies developed in this chapter can also be applied to reuse UL resources. We would only need to make a distinction between the two for cellular resource allocation since we assume half-duplex communications.

scenario. Although we study the simplified scenario, as illustrated in Fig. 6.1, the proposed framework and resource allocation methods in this paper are applicable to the general scenario with multiple UEs and FAPs, and will be discussed in their respective sections. A simple setup was also used in [102], with an included discussion on extending the model to more users.

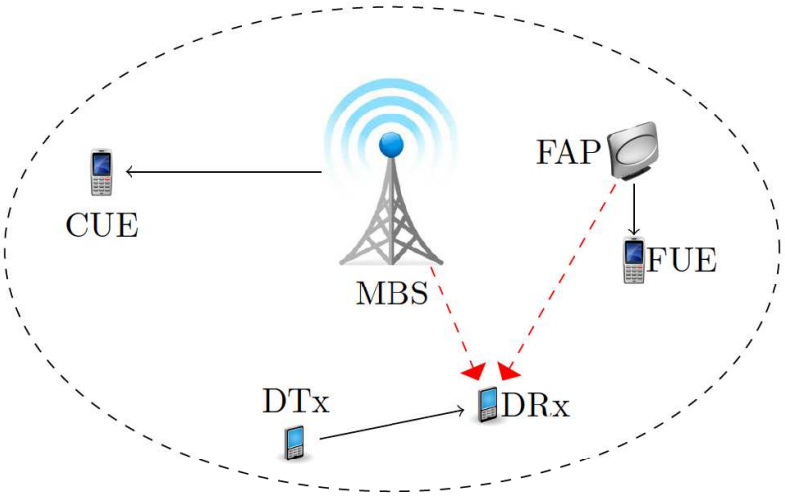

Figure 6.1: System model comprising of a D2D pair, MBS, FAP, and its served users. Strong interferences to the DRx from the MBS and FAP are shown in red dashed lines.

We assume that the MBS has perfect instantaneous channel state information (CSI) of all the links. This assumption has been widely used in the D2D literature [39, 40, 83, 84] and allows benchmark performance to be determined. The mode selection, resource allocation and power control is performed by the MBS in a centralized manner, based on the available perfect CSI. The transmit power of all transmitter nodes is denoted as P_t and the maximum transmit power is denoted as $P_t^{\max}$, where $t \in \{\mathrm{T}, \mathrm{M}, \mathrm{F}\}$ is the index for the transmitters, and T denotes DTx, M denotes MBS and F denotes FAP. The (minimum) rate at a receiver is denoted as ($\mathcal{R}_r^{\min}$) $\mathcal{R}_r$, while the corresponding (minimum) SINR under normalized resource allocation is denoted as ($\gamma_r^{\min}$) γ_r, where $r \in \{\mathrm{R}, \mathrm{C}, \mathrm{E}\}$ is the index for the receivers and R denotes DRx, C denotes CUE, and E denotes FUE. All the links are assumed to experience independent block fading.

The instantaneous channel coefficients are composed of small scale fading and

large scale path loss denoted as

$$g_{t,r} = h_{t,r} d_{t,r}^{-n} \tag{6.1}$$

where n is the path loss exponent, $h_{t,r}$ is the small scale Rayleigh fading coefficients, which are assumed to be independent and identically distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance and $d_{t,r}$ denotes the distance in meters between transmitter $t \in \{\mathrm{M}, \mathrm{T}, \mathrm{F}\}$ and receiver $r \in \{\mathrm{C}, \mathrm{R}, \mathrm{E}\}$. For simplicity, we denote the distance between DTx and DRx $d_{\mathrm{T,R}}$ as d. All links experience additive white Gaussian noise (AWGN) with power σ^2.

We use sum rate as our system performance metric with individual maximum power and minimum rate requirements. For clarity, due to the different nature of the D2D modes, we define each problem formulation in their respective sections.

6.2 Proposed Framework and Mode selection

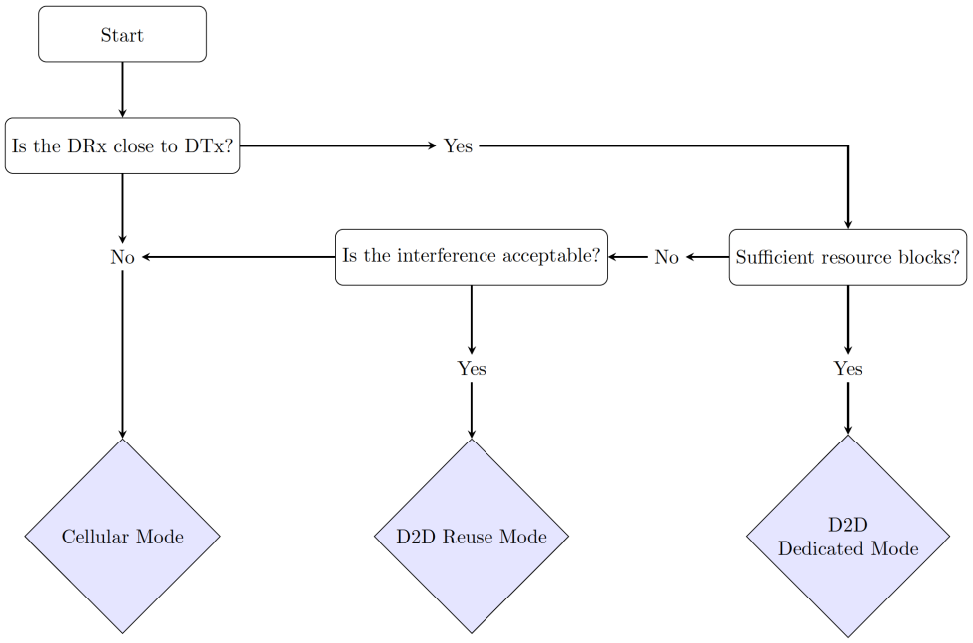

Figure 6.2: Proposed MBS assisted *D2D decision making framework* for mode selection, resource allocation and power control in D2D enabled two-tier cellular network.

We propose a base station assisted *D2D decision making framework*, as illustrated in Fig. 6.2, to enable the MBS to decide on the correct mode of D2D transmission, and determine the resource parameters of whichever mode is chosen (power for reuse mode, frequency resources for dedicated and cellular modes) that will maximize sum rate subject to maximum transmit power and minimum receiver rate constraints. The main steps in this process are described below:

1. The MBS first decides whether a potential D2D pair is close enough for D2D communications. Depending on the availability of orthogonal resources and potential interference, dedicated or reuse mode is chosen. Otherwise, the pair remain in cellular mode.

2. If the reuse mode is chosen, then the MBS instructs the CUE, DTx and FAP to control their transmit powers to guarantee quality of service to all receivers. This is done according to the approach proposed in Section 6.3.

3. If the cellular mode or the dedicated mode is chosen, then the MBS allocates resources to the CUE, FAP and D2D UEs. Since interference is not present, all transmitters can use the maximum transmit power. In the dedicated mode, we assume D2D UEs use the DL resources. In the cellular mode, we assume that both UL and DL resources are used by D2D UEs since the D2D communication is being relayed by the MBS.

6.2.1 Mode Selection

In order to allow as many potential D2D dedicated users as possible, our decision making framework firstly allows a potential D2D pair to enter dedicated mode if they are close enough and orthogonal resources are available. Interference is not considered in this case as using orthogonal resources eliminates interference. If orthogonal resources are not available, the decision to enter reuse mode is then made based on the potential interference. A potential D2D pair can only enter reuse mode if *both* of the following criteria are satisfied:

1. The DRx must be located outside an interference region such that the potential interference is lower than a threshold. Since actual powers are yet to be determined, we assume maximum transmit powers.

2. The distance d between the DTx and DRx must be less than a threshold (should be satisfied from initial step).

To determine the distance and interference thresholds, we recognize that in the cellular mode, the CUE and FUE should always experience better rates compared to D2D due to less interference. Therefore, mode selection is equivalent to finding under what conditions the DRx SINR in cellular mode is better than in D2D mode for the D2D users, i.e.,

$$\min\left(\frac{P_{\mathrm{T}}^{\max} g_{\mathrm{T,M}}}{P_{\mathrm{F}}^{\max} g_{\mathrm{F,M}} + \sigma^2}, \frac{P_{\mathrm{M}}^{\max} g_{\mathrm{M,R}}}{P_{\mathrm{F}}^{\max} g_{\mathrm{F,R}} + \sigma^2}\right) \geq \frac{P_{\mathrm{T}}^{\max} g_{\mathrm{T,R}}}{P_{\mathrm{M}}^{\max} g_{\mathrm{M,R}} + P_{\mathrm{F}}^{\max} g_{\mathrm{F,R}} + \sigma^2} \tag{6.2}$$

where the $\min(\cdot,\cdot)$ denotes the minimum operator and is used since the rate of the cellular two-hop link is limited by the minimum of the UL and the DL.

Suppose we consider a scenario where a D2D pair is close to each other, but located within a high interference region. From (6.2), if the interference is greater than a certain threshold

$$P_{\mathrm{M}}^{\max} g_{\mathrm{M,R}} + P_{\mathrm{F}}^{\max} g_{\mathrm{F,R}} \geq P_{\mathrm{T}}^{\max} g_{\mathrm{T,R}} \times \min\left(\frac{P_{\mathrm{F}}^{\max} g_{\mathrm{F,M}} + \sigma^2}{P_{\mathrm{T}}^{\max} g_{\mathrm{T,M}}}, \frac{P_{\mathrm{F}}^{\max} g_{\mathrm{F,R}} + \sigma^2}{P_{\mathrm{M}}^{\max} g_{\mathrm{M,R}}}\right) - \sigma^2, \tag{6.3}$$

using D2D mode will be an incorrect decision as it will lead to a lower rate. Note that this threshold value is a conservative estimate as it does not consider the overall system improvement from the CUE or FUE. When the correct mode selection decision is made, the rate achieved by the D2D pair, and also the overall system, will be greater.

A similar argument can be made for a second scenario where the DTx and DRx are located outside a high-interference region, but are far apart. Rearranging (6.2) to solve for the D2D separation distance, we find that D2D mode would be an incorrect decision if the DRx is outside the interference region, but the D2D pair is separated

by a distance of

$$d_{\text{adaptive}} \geq \sqrt[n]{\frac{h_{\text{T,R}} P_{\text{D}}^{\max}}{(P_{\text{M}}^{\max} g_{\text{M,R}} + P_{\text{F}}^{\max} g_{\text{F,R}} + \sigma^2)}} \times \sqrt[n]{\min\left(\frac{P_{\text{F}}^{\max} g_{\text{F,M}} + \sigma^2}{P_{\text{D}}^{\max} g_{\text{T,M}}}, \frac{P_{\text{F}}^{\max} g_{\text{F,R}} + \sigma^2}{P_{\text{M}}^{\max} g_{\text{M,R}}}\right)}. \tag{6.4}$$

where n is the path loss index.

However, when the DRx is close to an interference source, (6.4) may provide an unnecessarily small threshold, and therefore limit the number of D2D pairs. Thus, in our framework we chose the maximum threshold between (6.4) and a predetermined value d_{constant}, i.e.,

$$d \leq \max\{d_{\text{constant}}, d_{\text{adaptive}}\} \tag{6.5}$$

The benefits of this approach will be illustrated using simulation results in Section 6.5.1.

6.3 Power Allocation in Reuse Mode

In this section, we solve the overall sum throughput optimization problem in the reuse mode. In reuse mode, the problem reduces to finding the optimal powers that can maximize the sum throughput objective while meeting individual minimum rate requirements. We extend the method in [39] for the case of two transmitters to three transmitters to solve the power allocation problem in a two-tier cellular network. We present a geometric representation of the problem for the case of three transmitters (i.e., DTx, MBS and FAP) and present a near-optimal[4] solution approaching that of exhaustive search.

6.3.1 Problem Formulation

Our overall system aim is to maximize the sum rate with individual transmit power and receiver rate constraints. We can formulate the optimization problem as follows:

[4]We use the expression “near optimal" to describe the closeness of the solution to the optimal solution, rather than to define it as a specific solution or class of solutions.

$$\max_{P_{\mathrm{T}},P_{\mathrm{M}},P_{\mathrm{F}}} \left\{ \mathcal{R} \triangleq \log_2\left(1+\frac{P_{\mathrm{T}}g_{\mathrm{T,R}}}{P_{\mathrm{M}}g_{\mathrm{M,R}}+P_{\mathrm{F}}g_{\mathrm{F,R}}+\sigma^2}\right)+\log_2\left(1+\frac{P_{\mathrm{M}}g_{\mathrm{M,C}}}{P_{\mathrm{T}}g_{\mathrm{T,C}}+P_{\mathrm{F}}g_{\mathrm{F,C}}+\sigma^2}\right) + \log_2\left(1+\frac{P_{\mathrm{F}}g_{\mathrm{F,E}}}{P_{\mathrm{T}}g_{\mathrm{T,E}}+P_{\mathrm{M}}g_{\mathrm{M,E}}+\sigma^2}\right)\right\} \tag{6.6}$$

such that

$$P_t \leq P_t^{\max}, t \in \{\mathrm{M},\mathrm{T},\mathrm{F}\} \tag{6.7a}$$

$$\frac{P_{\mathrm{T}}g_{\mathrm{T,R}}}{P_{\mathrm{M}}g_{\mathrm{M,R}}+P_{\mathrm{F}}g_{\mathrm{F,R}}+\sigma^2} \geq \gamma_{\mathrm{R}}^{\min} \tag{6.7b}$$

$$\frac{P_{\mathrm{M}}g_{\mathrm{M,C}}}{P_{\mathrm{T}}g_{\mathrm{T,C}}+P_{\mathrm{F}}g_{\mathrm{F,C}}+\sigma^2} \geq \gamma_{\mathrm{C}}^{\min} \tag{6.7c}$$

$$\frac{P_{\mathrm{F}}g_{\mathrm{F,E}}}{P_{\mathrm{T}}g_{\mathrm{T,E}}+P_{\mathrm{M}}g_{\mathrm{M,E}}+\sigma^2} \geq \gamma_{\mathrm{E}}^{\min} \tag{6.7d}$$

where (6.7a) represents the maximum power constraints for each transmitter, while (6.7b)−(6.7d) are minimum SINR requirements. Note that for reuse mode, since all resources are shared and allocation is not considered, a minimum rate constraint is equivalent to a minimum SINR constraint.

6.3.2 Geometric Representation

We adopt a geometric approach to determine the optimal powers. To graphically represent the admissible powers, we first set orthogonal axes to be the powers. Next, setting constraints (6.7b)−(6.7d) to equality and rearranging, we obtain

$$f_{\mathrm{T}} \triangleq g_{\mathrm{T,R}}P_{\mathrm{T}} - \gamma_{\mathrm{R}}^{\min}g_{\mathrm{M,R}}P_{\mathrm{M}} - \gamma_{\mathrm{R}}^{\min}g_{\mathrm{F,R}}P_{\mathrm{F}} - \gamma_{\mathrm{R}}^{\min}\sigma^2 = 0, \tag{6.8a}$$

$$f_{\mathrm{M}} \triangleq -\gamma_{\mathrm{C}}^{\min}g_{\mathrm{T,C}}P_{\mathrm{T}} + g_{\mathrm{M,C}}P_{\mathrm{M}} - \gamma_{\mathrm{C}}^{\min}g_{\mathrm{F,C}}P_{\mathrm{F}} - \gamma_{\mathrm{C}}^{\min}\sigma^2 = 0, \tag{6.8b}$$

$$f_{\mathrm{F}} \triangleq -\gamma_{\mathrm{E}}^{\min}g_{\mathrm{T,E}}P_{\mathrm{T}} - \gamma_{\mathrm{E}}^{\min}g_{\mathrm{M,E}}P_{\mathrm{M}} + g_{\mathrm{F,E}}P_{\mathrm{F}} - \gamma_{\mathrm{E}}^{\min}\sigma^2 = 0, \tag{6.8c}$$

which represent planes in 3-dimensional space. The planes themselves represent the relationship between each node's power and the SINR thresholds. Each plane focuses on one threshold, and thus we refer to (6.8a)−(6.8c) as the D2D, MBS, and FAP

planes respectively. Each plane intersects with its respective axis at their respective minimum powers $P_t^{\min}$. Note that while the thresholds are stated in terms of the receiving node of that link, the powers are of the transmitting node.

We can plot (6.8a)−(6.8c) using their inequalities to obtain a 3-dimensional upper right corner region within a cube[5], the faces of which represent the maximum individual power constraints. The top right corner of this cube has the maximum power coordinates $(P_{\mathrm{T}}^{\max}, P_{\mathrm{M}}^{\max}, P_{\mathrm{F}}^{\max})$.

The smallest possible transmit powers, $P_t^{\min}$, that satisfy each users' SINR requirement can be calculated from (6.7b)−(6.7d) when there is no interference from the other transmissions. Therefore, the range of admissible powers is

$$P_{\mathrm{T}}^{\min} = \frac{\gamma_{\mathrm{R}}^{\min}\sigma^2}{g_{\mathrm{T,R}}} \leq P_{\mathrm{T}} \leq P_{\mathrm{T}}^{\max}, \tag{6.9a}$$

$$P_{\mathrm{M}}^{\min} = \frac{\gamma_{\mathrm{C}}^{\min}\sigma^2}{g_{\mathrm{M,C}}} \leq P_{\mathrm{M}} \leq P_{\mathrm{M}}^{\max}, \tag{6.9b}$$

$$P_{\mathrm{F}}^{\min} = \frac{\gamma_{\mathrm{E}}^{\min}\sigma^2}{g_{\mathrm{F,E}}} \leq P_{\mathrm{F}} \leq P_{\mathrm{F}}^{\max}. \tag{6.9c}$$

Meanwhile, the minimum powers that jointly satisfy the individual user rate constraints can be found by simultaneously solving (6.8a)−(6.8c) using standard methods such as Cramer's rule. Note that these powers will not maximize sum rate.

We assume that the coefficient matrix formed from (6.8a)−(6.8c) is full rank, i.e., the three planes intersect at a point Q, whose coordinates are all positive values since they represent transmission powers. Reuse mode is a viable option only if each signal strength is relatively large compared to the interference, making it easier to satisfy SINR constraints. This conclusion is consistent with others in the literature [32].

The admissible *power region* is formed by the intersection of the three planes in 3-dimensional space, and is bounded by these three planes and the three faces of the cube. The optimal powers lie within this power region. In order to avoid an extensive and inefficient exhaustive search, we propose a near-optimal solution which reduces the process to testing and selecting the candidate powers from a finite set.

[5]Strictly speaking, the region is a rectangular prism, but for conciseness we will use 'cube' to describe this region.

6.3.3 Proposed Solution - Vertex Search

In this paper, we adopt the simple approach of finding the corners or vertices of the power region to test for the optimal powers. This approach relies on the following two mathematical conditions:

1. The optimal powers cannot lie in the interior of the power region, and must be on a boundary.

2. The objective function is quasi-convex on a boundary, ensuring that the maximum values are at the endpoints/vertices.

The first condition was in fact proved in [88], and thus it is known that at least one of the powers is at its maximum when maximizing sum rate. However, this only states that the optimal solutions exist on the *boundary* of the power region, which includes vertices as well as higher dimensional edges and faces that contain an infinite number of points. Thus, this conclusion from [88] alone is not sufficient to obtain the finite set of points which will give the optimal solution. For two transmitters, it has been proven that the optimal power lies on the corners or vertices of the power region [39, 87], a fact that relies on the convexity of the sum rate function for two powers. However, it is well known that in general, the sum rate expression in (6.6) is non-convex with respect to arbitrary combinations of varying powers. Consequently, for arbitrary number of transmitters, the optimal powers may not necessarily lie on the vertices of the power region, leading to a possibly infinite set of points to test.

To prove the second condition and justify searching the vertices to maximize sum rate for arbitrary number of powers, we present the following two propositions.

Proposition 6.1. *Sum SINR is a quasi-convex function for any combination of varying powers. Hence, it is also jointly quasi-convex in all powers.*

Proof. See Appendix A.4. ■

Remark 6.1. Since sum SINR is a quasi-convex function, the powers maximizing it will lie on the one of the vertices of the power region.

Proposition 6.2. *When one receive power dominates, global maxima and minima for sum rate and sum SINR will occur at the same locations.*

Proof. We prove in Appendix A.5 that when one receive power dominates, e.g., an order of magnitude larger than others, sum SINR in (A.22) and the inner log term in (A.27) have the same *asymptotic* derivatives, meaning that the two functions will 'follow' each other more and more closely the larger the dominant power is. Since logarithm is a monotonic function and does not change the locations of local maxima or minima, this implies that the same powers that maximize sum SINR will also maximize sum rate. ■

Remark 6.2. Since global maxima of sum SINR will be at the vertices of the power region, the same vertices will also give near-optimal solutions for sum rate.

Note that approximations such as reformulating the objective function as a geometric program (GP) [103] can be used to solve (6.6). However, we show in the results section that our proposed simple approach yields near optimal solutions quite close to those obtained using exhaustive search and GP, but does not require an iterative approach.

A more detailed study of this gemetric approach for solving the power control problem is given in Appendix B, where we ask the question *Will the same set of powers that maximize sum SINR also maximize sum rate?* According to Proposition 6.2, sum rate and sum SINR will asymptotically have the same set of maximizing powers if one received power dominates the others.

6.3.4 Vertices of the Power Region

In this subsection, we present a systematic way of obtaining the coordinates of the vertices of the power region by solving relevant sets of SINR equations. All the vertex points are summarized in Table 6.1. The notation $\{P_a, P_b\}|^{\{f_a, f_b\}}$ means solve for powers P_a and P_b using simultaneous equations f_a and f_b with the other power maximized, where $a, b \in \{\mathrm{T}, \mathrm{M}, \mathrm{F}\}$.

Face points with one power maximized: There exists vertices that lie on a face of the cube and are formed from the intersection of two planes, e.g., point F in

Table 6.1: Finite set of vertices (suboptimal powers) for reuse mode.

Type	Condition	Number of points	Set of vertices (suboptimal powers)
Face point	All	9	$(\{P_T, P_M\}\|^{\{f_T,f_M\}}, P_F^{max})$, $(\{P_T, P_M\}\|^{\{f_T,f_F\}}, P_F^{max})$, $(\{P_T, P_M\}\|^{\{f_M,f_F\}}, P_F^{max})$, $(\{P_T, P_F\}\|^{\{f_T,f_M\}}, P_M^{max})$, $(\{P_T, P_F\}\|^{\{f_T,f_F\}}, P_M^{max})$, $(\{P_T, P_F\}\|^{\{f_M,f_F\}}, P_M^{max})$, $(P_T^{max}, \{P_M, P_F\}\|^{\{f_T,f_M\}})$,$(P_T^{max}, \{P_M, P_F\}\|^{\{f_T,f_F\}})$, $(P_T^{max}, \{P_M, P_F\}\|^{\{f_M,f_F\}})$
Edge point - All thresholds satisfied	$\gamma'_R \geq \gamma_R^{min}, \gamma'_C \geq \gamma_C^{min}$ and $\gamma'_E \geq \gamma_E^{min}$	4	$(P_T\|^{\{f_T\}}, P_M^{max}, P_F^{max})$, $(P_T^{max}, P_M\|^{\{f_M\}}, P_F^{max})$, $(P_T^{max}, P_M^{max}, P_F\|^{\{f_F\}})$, $(P_T^{max}, P_M^{max}, P_F^{max})$
Edge point - Two thresholds satisfied	$\gamma'_E \leq \gamma_E^{min}$	4	$(P_T\|^{\{f_T\}}, P_M^{max}, P_F^{max})$, $(P_T\|^{\{f_F\}}, P_M^{max}, P_F^{max})$ $(P_T^{max}, P_M\|^{\{f_M\}}, P_F^{max})$, $(P_T^{max}, P_M\|^{\{f_F\}}, P_F^{max})$
	$\gamma'_R \leq \gamma_R^{min}$	4	$(P_T^{max}, P_M\|^{\{f_T\}}, P_F^{max})$, $(P_T^{max}, P_M\|^{\{f_M\}}, P_F^{max})$ $(P_T^{max}, P_M^{max}, P_F\|^{\{f_T\}})$, $(P_T^{max}, P_M^{max}, P_F\|^{\{f_F\}})$
	$\gamma'_C \leq \gamma_C^{min}$	4	$(P_T\|^{\{f_T\}}, P_M^{max}, P_F^{max})$, $(P_T\|^{\{f_M\}}, P_M^{max}, P_F^{max})$ $(P_T^{max}, P_M^{max}, P_F\|^{\{f_M\}})$, $(P_T^{max}, P_M^{max}, P_F\|^{\{f_F\}})$
Edge point - One threshold satisfied	$\gamma'_E \leq \gamma_E^{min}$ and $\gamma'_C \leq \gamma_C^{min}$	2	$(P_T\|^{\{f_M\}}, P_M^{max}, P_F^{max})$, $(P_T\|^{\{f_F\}}, P_M^{max}, P_F^{max})$
	$\gamma'_R \leq \gamma_R^{min}$ and $\gamma'_E \leq \gamma_E^{min}$	2	$(P_T^{max}, P_M\|^{\{f_T\}}, P_F^{max})$, $(P_T^{max}, P_M\|^{\{f_F\}}, P_F^{max})$
	$\gamma'_R \leq \gamma_R^{min}$ and $\gamma'_C \leq \gamma_C^{min}$	2	$(P_T^{max}, P_M^{max}, P_F\|^{\{f_T\}})$, $(P_T^{max}, P_M^{max}, P_F\|^{\{f_M\}})$

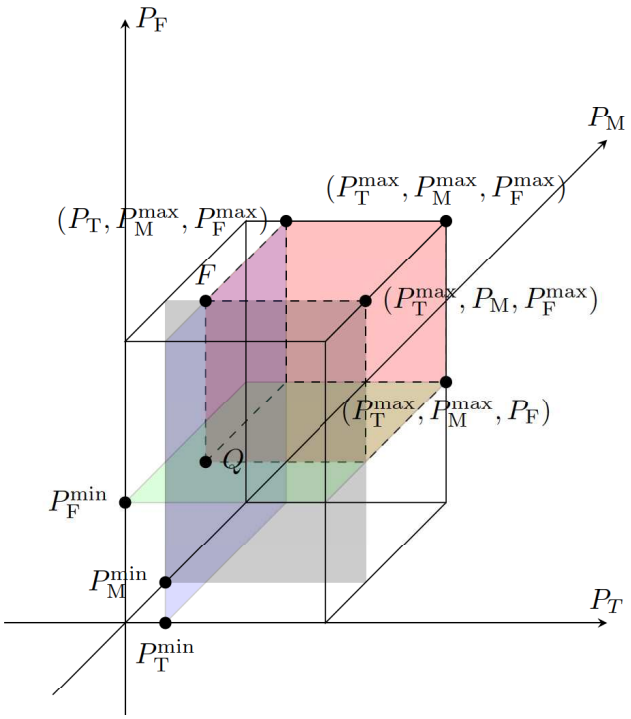

Figure 6.3: All thresholds are satisfied.

Fig. 6.3. There are nine such vertices (three faces with three ways of choosing two intersecting planes for each face). These vertices can be found by solving two plane equations simultaneously with the power corresponding to the third face maximized. In general, it is difficult to identify exactly which of these nine points may be optimal for a given interference scenario. Thus, we need to test all nine vertices.

Edge points with maximum powers satisfying all thresholds: Consider the case where the three planes are orthogonal, as shown in Fig. 6.3. In this case, the power region includes the top corner of the cube, where all three powers are maximized, and three other corner points where the planes intersect the edges of the cube, which we shall label as *edge points*. Since the top corner lies in the power region, this indicates that when all powers are maximized, all three SINRs γ_r are greater than their minimum thresholds $\gamma_r^{\min}$. For the rest of this section, we denote γ_r' as the SINR for each node when all powers are at their maximum. There are four such points, as summarized in Table I. Note that the same SINR scenario can occur even

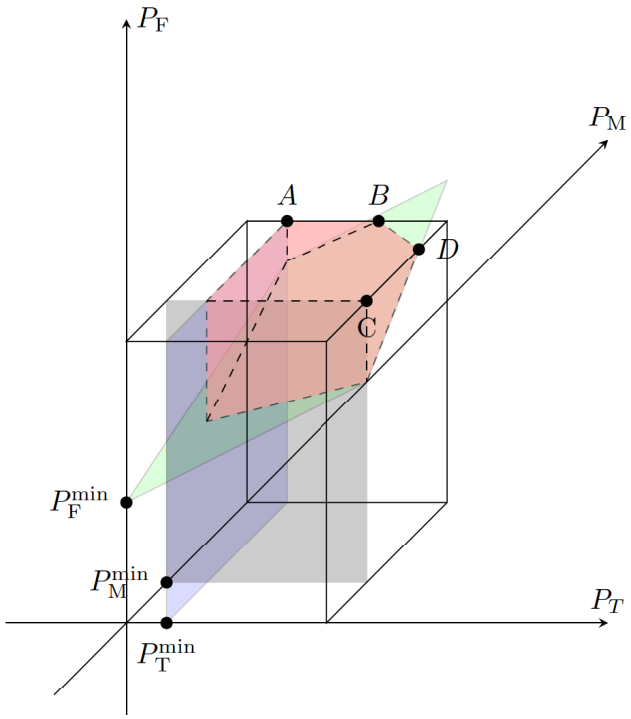

Figure 6.4: Two thresholds are satisfied.

when the planes are not perpendicular.[6] The distinctive feature of this scenario is that the top corner is within the region spanned by the planes, and that each plane only intersects one of the maximum power edges of the cube.

Edge points with maximum powers satisfying two thresholds: To visualize this scenario, imagine tilting the planes pivoted at Q to form new power regions. For instance, if we tilt only the FAP plane upwards, it will eventually pass through the top corner and intersect the other two top edges. These two additional edge points (B and D in Fig. 6.4) add to the existing two edge points (A and C) to give a total of four edge points on the cube's edges. Since the top corner point will now be below the FAP plane, this means that $\gamma_{\mathrm{E}}' \leq \gamma_{\mathrm{E}}^{\min}$. Similar arguments can be made for the other two planes, giving us three cases where there are a total of four corner points in the power region, each case corresponding to one γ_r' that is less than its respective threshold.

[6] In fact, perpendicular planes which each only intersect one axis corresponds to an interference-free scenario.

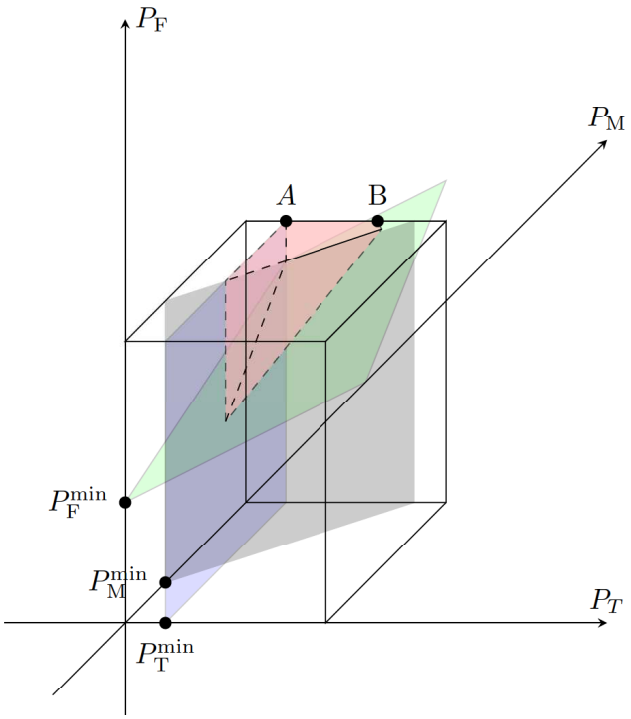

Figure 6.5: One threshold is satisfied.

Edge points with maximum powers satisfying one threshold: For scenarios where two γ_r' fail to reach their thresholds and only one is met, the two planes will be tilted such that the corner point lies outside both their feasible regions, as shown in Fig. 6.5 where the FAP and MBS planes lie above and to the left of the top corner respectively. In these cases, the feasible region will intersect one of the three corner edges at two points. Fig. 6.5 illustrates the maximum MBS power edge being intersected at two points A and B by the D2D and FAP planes respectively. Note that the MBS plane also intersects the same edge, but that point of intersection is outside the power region. Thus, we get two points, each corresponding to a set of conditions.

6.4 Resource Allocation in Dedicated and Cellular Modes

If mode selection decides that the D2D pair can transmit using either dedicated or cellular mode, time and/or frequency resources must be allocated. We make the following assumptions for resource sharing in both dedicated D2D and cellular mode:

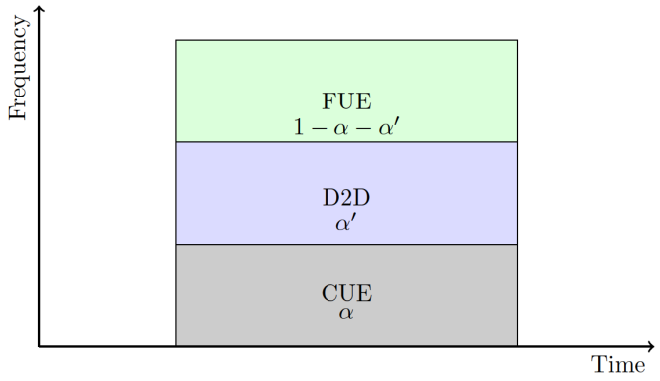

Figure 6.6: Frequency sharing in dedicated mode.

(i) since cellular frequencies are used, there is a minimum rate guarantee for each user, including the DRx, (ii) there are enough resources to meet all users' minimum rate requirements, and (iii) at any one time, one transmitter can only operate in either uplink or downlink, i.e., half duplex.

6.4.1 Problem Formulation

Since there is no interference in both dedicated and cellular modes and all powers can be maximized, the SINR at each receiver is the same as the signal-to-noise ratio (SNR) at that receiver, given as[7]

$$\gamma_{\mathrm{C}} = \frac{g_{\mathrm{M,C}} P_{\mathrm{M}}^{\max}}{\sigma^2}, \qquad \gamma_{\mathrm{R}} = \frac{g_{\mathrm{T,R}} P_{\mathrm{T}}^{\max}}{\sigma^2}, \qquad \gamma_{\mathrm{E}} = \frac{g_{\mathrm{F,E}} P_{\mathrm{F}}^{\max}}{\sigma^2}. \tag{6.10}$$

We formulate a general optimization for a Long Term Evolution (LTE)-like resource grid with distinct resource blocks as follows

$$\underset{B_r^i}{\text{maximize}} \quad \sum_r \sum_i^{N^t} B_r^i \delta^f \delta^t \log_2 \left(1 + \frac{\gamma_r}{B_r^i \delta^f}\right) \tag{6.11}$$

$$\text{subject to} \quad \sum_i^{N^t} B_r^i \delta^f \delta^t \log_2 \left(1 + \frac{\gamma_r}{B_r^i \delta^f}\right) \geq \mathcal{R}_r^{min} \tag{6.12}$$

[7]With slight abuse of notation but for the sake of simplicity, we use the symbol γ_r for SNR, where as in Section II we denoted minimum SINR at a receiver as $\gamma_r^{\min}$, where $r \in \{\mathrm{R}, \mathrm{C}, \mathrm{E}\}$ is the index for the receivers.

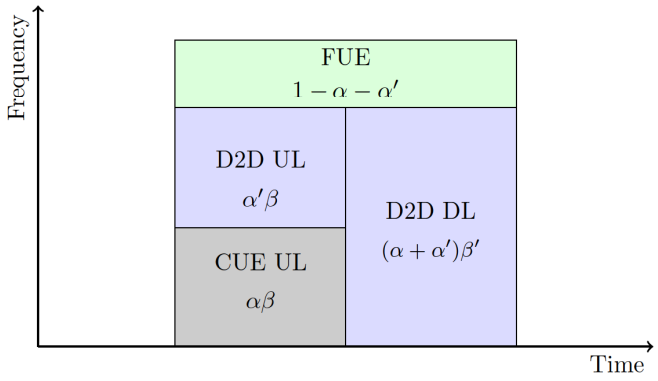

Figure 6.7: Frequency sharing in cellular mode.

where B_r^i is the number of resource blocks for user r at the ith time interval, δ_r^f and δ_r^t are the (constant) fractions representing the portion of each frequency and time block compared to the total grid respectively, and N^t is the total number of time intervals for the resource grid. The total number of blocks allocated to user r is therefore $\sum_i^{N^t} B_r^i$. Note that we divide the SNR by the frequency portion as we define σ^2 with respect to the entire bandwidth, resulting in equal noise power density [85].

The general formulation is difficult to solve, and in practice requires numerical methods. In order to gain insight into generalizations for arbitrary number of users and to obtain closed form solutions, we can show that for a given $\sum_i^{N^t} B_r^i = \mathcal{N}$ number of resource blocks and assuming that each block is no more preferable to any other, allocating resources across frequency will produce higher rates than allocating across time or in a random manner. For ease of proof and without loss of generality, we assume high SNR such that $\log_2(1+SNR) \approx \log_2(SNR)$. Using $\log_2(A)+\log_2(B) = \log_2(AB)$ and the arithmetic-geometric mean inequality, which states that the maximum of a product of terms with a sum constraint occurs when all terms are equal, we find that the maximum of

$$\sum_i^{N^t} B_r^i \log_2\left(\frac{\gamma_r}{B_r^i}\right) = \log_2 \prod_i^{N^t} \left(\frac{\gamma_r}{B_r^i}\right)^{B_r^i} \tag{6.13}$$

$$\text{subject to} \quad \sum_i^{N^t} B_r^i = \mathcal{N} \tag{6.14}$$

occurs when all B_r^i are equal. In other words, using equal bandwidth allocation across all time intervals for each user will provide the largest rates. Thus, although we can generalize resource allocation to be compatible with arbitrary allocations, frequency allocation will give higher rates compared to other approaches for a given number of resources blocks.

Since frequency allocation can be solved in closed form, we analyze frequency allocation formulated as follows:

$$\underset{x_r}{\text{maximize}} \qquad \sum x_r \log_2\left(1+\frac{\gamma_r}{x_r}\right) \tag{6.15a}$$

$$\text{subject to} \qquad x_r \log_2\left(1+\frac{\gamma_r}{x_r}\right) \geq \mathcal{R}_r^{min} \tag{6.15b}$$

where the factor x_r is a function of the resource portions $0 \leq \alpha, \alpha', \beta, \beta' \leq 1$. For ease of analysis, we study only the case where an exact bandwidth is allocated across all time intervals.

6.4.2 Frequency Sharing in Dedicated D2D Mode

For the allocation structure illustrated in Fig. 6.6, we want to maximize

$$\mathcal{R}_2 = \alpha \log_2\left(1+\frac{\gamma_{\mathrm{C}}}{\alpha}\right) + \alpha' \log_2\left(1+\frac{\gamma_{\mathrm{R}}}{\alpha'}\right) + (1-\alpha-\alpha') \log_2\left(1+\frac{\gamma_{\mathrm{E}}}{1-\alpha-\alpha'}\right). \tag{6.16}$$

6.4.2.1 Unconstrained

With no minimum rate constraints, we differentiate (6.16) with respect to α and α', and simultaneously solve for $\frac{\partial \mathcal{R}_2}{\partial \alpha} = 0$ and $\frac{\partial \mathcal{R}_2}{\partial \alpha'} = 0$, which gives us the solutions

$$\alpha = \frac{\gamma_{\mathrm{C}}}{\gamma_{\mathrm{C}} + \gamma_{\mathrm{R}} + \gamma_{\mathrm{E}}}, \tag{6.17a}$$

$$\alpha' = \frac{\gamma_{\mathrm{R}}}{\gamma_{\mathrm{C}} + \gamma_{\mathrm{R}} + \gamma_{\mathrm{E}}}. \tag{6.17b}$$

Substituting the above into (6.16) and simplifying, the optimal sum rate is

$$\mathcal{R}_2^{\mathrm{opt}} = \log_2(1 + \gamma_{\mathrm{C}} + \gamma_{\mathrm{R}} + \gamma_{\mathrm{E}}). \tag{6.18}$$

6.4.2.2 Constrained

To meet each user's minimum rate requirement, we require the solution to

$$\alpha \log_2 \left(1 + \frac{\gamma_r}{\alpha}\right) = \mathcal{R}_r^{\min} \tag{6.19}$$

for each user. The solution can be written in terms of the Lambert W function (see Appendix A.7), but there is no analytical solution that can be expressed using elementary functions. A simple numerical line search along $0 \leq \alpha \leq 1$ can be used to find an optimal solution.

6.4.3 Frequency Sharing in Cellular D2D Mode

In frequency sharing cellular mode, because there is only one MBS transmitter, the D2D UL and CUE UL must occur at the same time, with D2D DL occurring immediately afterwards. The FUE can be allocated subbands at any time as it is served by a separate transmitter. Therefore, the allocation scheme follows the partitions as shown in Fig. 6.7.

In frequency sharing in cellular mode, the sum rate to be optimized is

$$\begin{aligned}\mathcal{R}_4 = {} & \alpha\beta \log_2 \left(1 + \frac{\gamma_{\mathrm{M,UL}}}{\alpha}\right) + \min\left(\alpha'\beta \log_2 \left(1 + \frac{\gamma_{\mathrm{R,UL}}}{\alpha'}\right), (\alpha + \alpha')\beta' \log_2 \left(1 + \frac{\gamma_{\mathrm{R,DL}}}{\alpha + \alpha'}\right)\right) \\ & + (1 - \alpha - \alpha') \log_2 \left(1 + \frac{\gamma_{\mathrm{E}}}{(1 - \alpha - \alpha')}\right),\end{aligned} \tag{6.20}$$

where $\beta + \beta' = 1$[8] and $\gamma_{r,\mathrm{UL}} = \frac{g_{\mathrm{C,M}} P_{\mathrm{C}}^{\max}}{\sigma^2}$ is the SNR at the MBS during CUE UL with the CUE transmitting at its maximum power $P_{\mathrm{C}}^{\max}$.

[8]Setting $\beta' = 0$ would be equivalent to having two cellular users, and the solution would be the same as that in Section V-B.

6.4.3.1 Unconstrained

We define uplink and downlink rates as

$$\mathcal{R}_{\mathrm{UL}} = \log_2\left(1 + \frac{\gamma_{\mathrm{R,UL}}}{\alpha'}\right), \tag{6.21a}$$

$$\mathcal{R}_{\mathrm{DL}} = \log_2\left(1 + \frac{\gamma_{\mathrm{R,DL}}}{\alpha + \alpha'}\right). \tag{6.21b}$$

To simplify (6.20) into an expression involving only α and α', we note that the maximum sum rate occurs when $\alpha'\beta\mathcal{R}_{\mathrm{UL}} = (\alpha + \alpha')\beta'\mathcal{R}_{\mathrm{DL}}$, with the solution given by

$$\beta = \frac{\mathcal{R}_{\mathrm{DL}}}{\frac{\alpha'}{\alpha+\alpha'}\mathcal{R}_{\mathrm{UL}} + \mathcal{R}_{\mathrm{DL}}}. \tag{6.22}$$

Substituting the above, the rate expression for D2D is

$$\mathcal{R}_d(\alpha) = \frac{\alpha'\mathcal{R}_{\mathrm{UL}}\mathcal{R}_{\mathrm{DL}}}{\frac{\alpha'}{\alpha+\alpha'}\mathcal{R}_{\mathrm{UL}} + \mathcal{R}_{\mathrm{DL}}}. \tag{6.23}$$

Therefore, we can simplify (6.20) to

$$\mathcal{R}_4 = \frac{\alpha\mathcal{R}_{\mathrm{DL}}\log_2(1 + \frac{\gamma_{\mathrm{R,UL}}}{\alpha})}{\frac{\alpha'}{\alpha+\alpha'}\mathcal{R}_{\mathrm{UL}} + \mathcal{R}_{\mathrm{DL}}} + \mathcal{R}_d(\alpha) + (1 - \alpha - \alpha')\log_2\left(1 + \frac{\gamma_{\mathrm{E}}}{(1 - \alpha - \alpha')}\right). \tag{6.24}$$

A numerical search for $0 \leq \alpha + \alpha' \leq 1$ can be performed to maximize (6.24).

6.4.3.2 Constrained

In this scenario, we desire to maximize (6.24) under a minimum rate constraint for each user. Again, a numerical search can be performed to find the maximum.

6.5 Results and Discussion

In this section, we present simulation results to illustrate the benefits of using our decision making framework over conventional cellular transmission for a potential D2D pair. Unless stated otherwise, simulation parameters presented in Table 6.2 are

Table 6.2: Values of Simulation Parameters

Parameter	Value
Bandwidth	20 MHz
Noise spectral density	-174 dBm/Hz
Max MBS transmit power	$P_{\mathrm{M}}^{\max} = 43$ dBm
Max FAP transmit power	$P_{\mathrm{F}}^{\max} = 21$ dBm
Max DTx transmit power	$P_{\mathrm{T}}^{\max} = 23$ dBm
DTx coordinates	Varying along $x = y$
DRx coordinates	Varying along $x = y$
MBS coordinates	(0, 0)
CUE coordinates	(500, 0)
FAP coordinates	(100, 200)
FUE coordinates	(110, 200)
DTx to DRx pathloss	$28 + 40\log_{10}(d)$ (dB)
MBS to CUE pathloss	$15.3 + 37.6\log_{10}(d_{\mathrm{M,C}})$ (dB)
FAP to FUE pathloss	$38.5 + 20\log_{10}(d_{F,E})$ (dB)
CUE minimum SINR	$\gamma_{\mathrm{C}}^{\min} = 0$ dB
FUE minimum SINR	$\gamma_{\mathrm{E}}^{\min} = 7$ dB
DRx minimum SINR	$\gamma_{\mathrm{R}}^{\min} = 3$ dB

used, which are similar to those adopted in [40]. We use (x, y) coordinates in meters to describe node locations.

6.5.1 Mode Selection

We first show the advantages of using (6.5) compared to using a constant and adaptive distance threshold only. Setting $d_{\text{constant}} = 50$m and assuming orthogonal resources are available 50% of the time, Fig. 6.8 shows that picking the largest threshold between the predetermined and calculated gives the highest percentages of users entering dedicated mode. When interference to the DRx is large, choosing a predetermined distance threshold is more beneficial. When the DRx is farther from an interference source, an adaptive threshold is the better choice as larger D2D separation distances can be tolerated. It is evident that the proposed method captures the best features of the other two, and in fact slightly outperforms the best of both at every location tested.

When orthogonal resources are not available, Fig. 6.9 plots the D2D rate gain versus the distance between the DTx and DRx, d. The D2D rate gain refers to the ratio between the D2D rate and the cellular rate, both under the same interference condi-

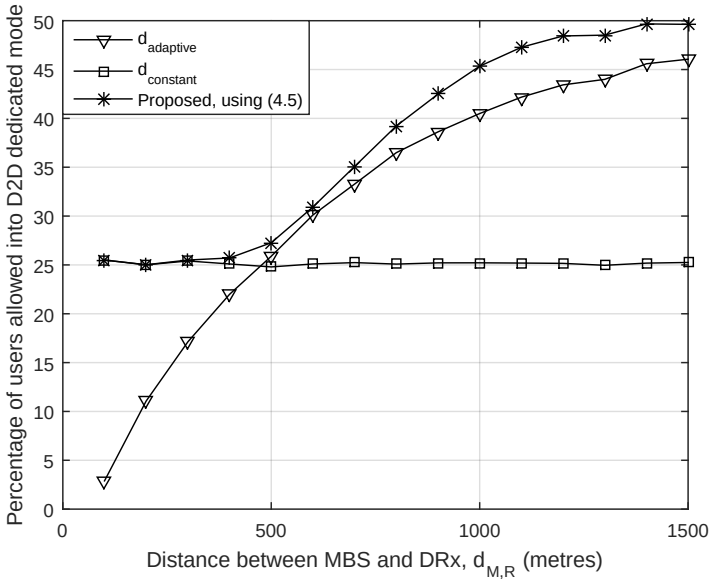

Figure 6.8: Percentage of potential D2D pairs entering dedicated mode. Predetermined threshold is better when interference is large, while adaptive threshold is better when interference is small.

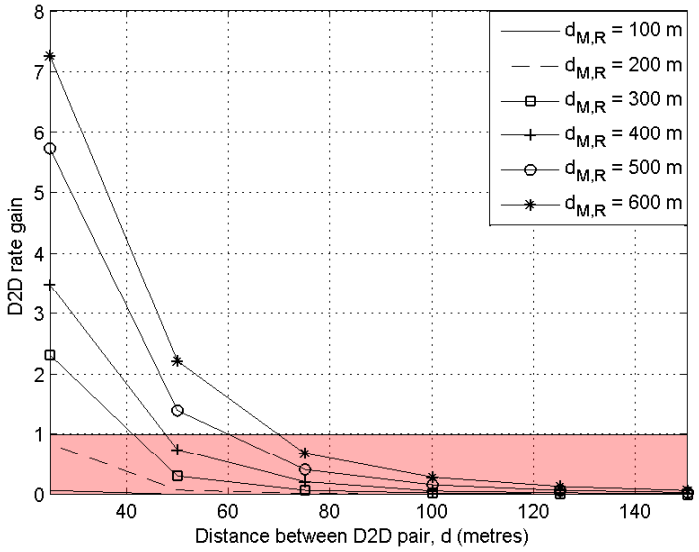

Figure 6.9: D2D rate gain versus the distance between the DTx and DRx, d for different MBS-DRx distance, $d_{\mathrm{M,R}}$. The shaded area below D2D rate gain of 1 represents the region where selecting D2D mode would be an incorrect decision.

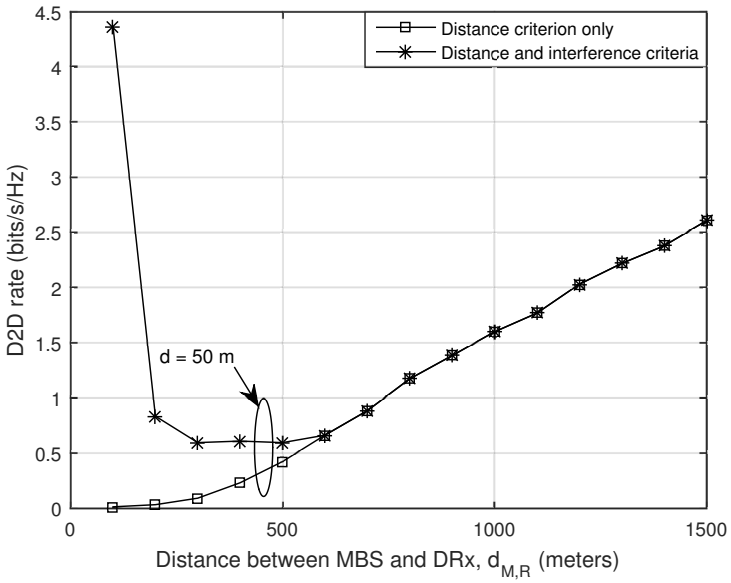

Figure 6.10: D2D rate versus the distance between the MBS and DRx, $d_{\mathrm{M,R}}$, for mode selection using distance only criterion and two stage criteria.

tions. We can see that the D2D gain decreases when the D2D pair become farther apart and also when the DRx is closer to the MBS. This is in line with the discussion in Section 6.2 since: (i) when the DRx is closer to the MBS (the largest interference source), using cellular mode should provide higher rates than an incorrect D2D mode decision since there would have been more interference, and (ii) when the D2D pair separation distance increases, cellular mode should provide higher rates since D2D mode would be weaker with increasing separation distance under constant transmit power. In Fig. 6.9, the D2D separation distance at which each curve intersects the boundary of this region can be calculated using (6.4). Our calculated and simulated values were found to be in close agreement. For example, the calculated separation distance for $d_{\mathrm{M,R}} = 600$ m is 75.9 m, while the simulations give a value of 71 m.

Fig. 6.10 shows the actual rates experienced by a DRx when using the proposed mode selection method satisfying (6.2) and when using just the D2D minimum distance criterion with $d = 50$ m. If D2D mode is always allowed for $d = 50$ m, the DRx can experience a smaller rate due to its close proximity to an MBS (or other large interference source), while using our proposed method will avoid such instances.

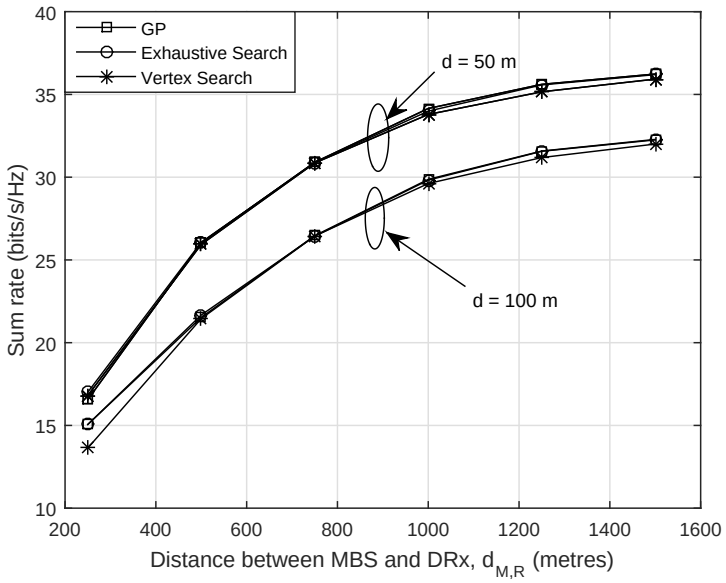

Figure 6.11: Sum rate in reuse mode with transmit powers determined using proposed near-optimal vertex search approach, geometric programming and exhaustive search.

It is important to note that our results in this subsection do not suggest that cellular mode is superior to D2D mode. Rather, our results highlight that under some conditions, using a single criterion to determine mode selection can lead to an incorrect decision. We will show in Section 6.5.3 that if D2D is operating in dedicated mode, it can outperform cellular mode.

6.5.2 Reuse Mode

Fig. 6.11 plots the sum rate in reuse mode versus varying MBS-DRx distance, $d_{\mathrm{M,R}}$, comparing the near-optimal powers found using the proposed approach with those obtained from geometric programming (GP) [103] and exhaustive search. We believe that using GP and exhaustive search serve as sufficient benchmarks - GP is one of the most common numerical approaches to finding near optimal solutions for the power control problem, while exhaustive search with a sufficiently fine step size confirms the optimality.

Our results show that for the considered parameters, our proposed method of searching the vertices of the power region and using the one that gives the maximum

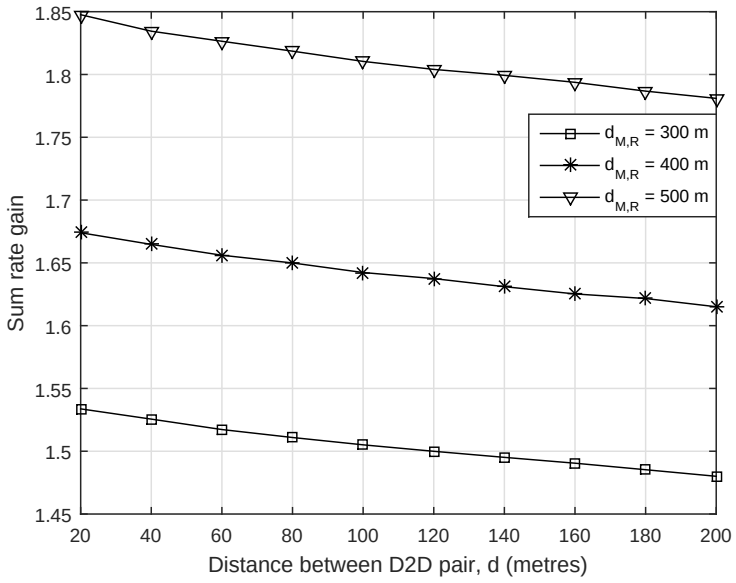

Figure 6.12: Sum rate gain versus the distance between the DTx and DRx, d for constrained frequency resource sharing.

sum rate is comparable to optimal solutions. However, our method requires far fewer calculations than exhaustive search and GP, the latter of which relies on successive approximations with no prior indication on how many iterations are required. For example, using an Intel i7 3.2 GHz CPU with 16 GB RAM, for $d_{\mathrm{M,R}} = 1000$ m in Fig. 6.11 GP took up to 44.5 seconds to calculate a solution, exhaustive search took 49.8 seconds, while our vertex search took only 1.2 seconds, i.e., an improvement of around 40 times over both benchmarks. Thus, GP can be unreliable in determining a suboptimal solution in sufficient time, while our vertex search approach will always return a suboptimal solution if the problem is feasible for small numbers of reuse powers. A further advantage of vertex search is that it always takes approximately the same time to calculate a solution for each realization, while the run time and accuracy of GP heavily depends on stoppage parameters.

6.5.3 Dedicated and Cellular Modes

Fig. 6.12 shows the sum rate gain, i.e., sum rate in dedicated mode divided by sum rate in cellular mode, under minimum rate requirements for each user. It is clear that

dedicated D2D mode provides a greater sum rate when the D2D separation distance is small, and/or when the MBS-DRx distance is large.

It must be noted that the unconstrained dedicated and cellular modes offered similar sum rates under the simulation parameters, and thus their results are not shown. It is clear however that unconstrained dedicated sum rates will never be lower than their cellular counterparts since the D2D option is intended to improve overall system performance, and will not degrade the best performing user. Thus, we can conclude that D2D mode is more advantageous when users have individual rate constraints.

6.5.4 Scalability Discussion

Although we have presented our methodology using a simple system model, we can analyze the scalability with respect to increasing number of base stations and users. For our mode selection framework, the number of decisions scales linearly with the number of potential D2D pairs, and not the total number of users or base stations.

For reuse mode, we have presented three transmit powers to be optimized, leading to a 3-dimensional problem. Increasing the number of transmit powers increases the dimensionality of the problem, while increasing the number of users increases the number of planes that restrict the size of the power region. For N powers, the power constraints form an N-dimensional hypercube, while the minimum SINR constraints further bound the region to form an N-dimensional polytope. Depending on which scenario the network is in (i.e., number of thresholds satisfied by max powers), the complexity of vertex search could increase exponentially at worst (e.g., in Fig. 6.3), and linearly at best (Fig. 6.5). However, if a small number of users share the same resource, since our vertex search avoids iterations, it can still be a more effective solution. Exact expressions for the vertices for N-dimensions is an interesting topic for future research.

For frequency sharing in dedicated mode (illustrated in Fig. 6.6), we prove in Appendix A.6 that the unconstrained case has a general solution for any number of transmitters and partitions. The general constrained case also has a solution given

by solving (6.19) for each user. With increasing number of nodes (base station or user), the complexity would scale linearly as each additional node would require one additional equation to solve for (from differentiating (6.16) or solving (6.19)). Frequency sharing in cellular mode (illustrated in Fig. 6.7) can have various allocation structures due to the simultaneous uplink condition, and thus cannot be generalized in the present manner.

6.6 Summary

We have presented a comprehensive mode selection, power control and resource allocation framework for D2D communication underlaying a two-tier cellular network. Our proposed mode selection scheme allows D2D communications under stricter conditions, leading to more correct decision making and a higher rate of allowing dedicated mode. We have also proposed a geometric approach to determine near-optimal powers for power allocation in reuse mode with faster computational time than benchmark methods, and provided closed-form resource allocations for orthogonal D2D mode for any number of users.

Chapter 7

Conclusions

This thesis has investigated problems, proposed solutions and provided insights into four main features of HetNets that will enable the growth and evolution of 5G cellular networks.

The first half of this thesis studied two problems that arise from having large differences in base station transmit power in a HetNet. Chapter 2 proposed a precoder with a generalized inverse matrix structure to suppress or eliminate interference from a macro base station to an external femto user. Our results showed that without power constraints, it is possible to completely eliminate interference if perfect CSI is available. More practically, we showed that using Fourier series estimates in the case of imperfect CSI, or a slight compromise in average macro user rates if a power constraint exists, can still achieve satisfactory interference suppression.

Chapter 3 studied the effects of using a dynamic bias function over a static bias value. We illustrated two association orders, and derived equivalent radii and association probabilities to show the benefits and properties of dynamic biasing. Our results found that dynamic biasing by associating closest users first acts as a natural prevention against small cell overloading.

We presented a discussion on the differences between balance and fairness in Chapter 4, and introduced our notion of expected network load that is independent of user distribution. We proposed a network balance index which is more useful in improving sum rate for clustered networks than fairness is. Ultimately, our view is that network balance is similar but a distinct concept to user fairness, and can provide more useful information about a network in certain scenarios.

The second half of this thesis studied advanced HetNet concepts beyond small

cells. Chapter 5 discussed the behaviour of network associations with network dynamics. We first introduced a base station preference association scheme where users associate with the base stations where they are most preferred, regardless of actual received powers or potential rates. We proved analytically and via simulations that this association leads to high user fairness. Next, we studied how associations will change with entering or exiting users and base stations, and found that the users who are most likely to switch associations are those who have average associated ranks. We also verified that a shrinking network has more effect on user association than a growing one.

Chapter 6 proposed a D2D decision making framework that determines mode selection using both an interference and distance criteria, then computes suitable power and resource parameters. We analysed resource allocation for both orthogonal and non-orthogonal modes. A near-optimal method which significantly reduces the search set was proposed for power control, while for orthogonal modes we showed that D2D is more effective when users have individual rate constraints.

Overall, this thesis has provided insight into better network decision making to enhance individual and total network rates, improve resource management, and reduce computation for user association in 5G HetNets.

7.1 Future Research Directions

The field of HetNets is a vastly rich research area with tremendous potential. The following major research directions may be the focus of future work:

Interference: Base station cooperation could be introduced for interference management to reduce interference for a larger number of users, or improve the effectiveness by providing more accurate network and channel information through base station cooperation [19, 104]. Precoding methods could be combined with multiple access techniques to form a comprehensive interference management strategy for varying scenarios and user behaviours.

Network Evolution: An exciting future work could be using Markov decision processes and game theory to model network dynamics and predict future states. To

extend this idea, instead of individual users or base stations, correlated behaviour such as groups of users entering or leaving (e.g., commuters getting off a train station) can be modelled.

User Association: Most studies on user association have been to maximize one objective (usually sum rate or some variant of), with some constraint on transmit power, fairness or some secondary metric. A difficult but interesting research direction is use multiobjective optimization, where multiple objectives are jointly optimized. Such a formulation is difficult to solve, and requires careful treatment of all objective functions, e.g. through selection of objective weights. Pareto optimal solutions might be computationally extensive, but practical sub-optimal algorithms are possible alternatives.

Massive MIMO and mmWaves: Two other key innovations for 5G network are massive MIMO and mmWaves, which can both be integrated into HetNet designs [7, 21, 105] and are feasible to implement [106]. mmWaves will require small antennas sizes, which are suitable for small cells or even mobile devices. Meanwhile, massive MIMO can replace conventional macro base stations, further justifying the importance of small cell load balancing. The harmony of these three innovations will provide the means to support future wireless demands both for the fronthaul and backhaul networks.

Asymmetry: The vast differences in transmit powers of macro base stations and small cells means that downlink and uplink communications can be very different [107]. For instance, to reduce power usage at a user mobile device, downlink communications can be sent from a macro base station, but uplink communications can be received by a nearby small cell. Such a concept leads to different association schemes, as well as more complex resource management.

Mobility and Handover: Smaller coverage areas of small cells means that highly mobile users will move in and out of service regions rapidly, leading to mobility and handover issues [2]. Some preliminary research have been done on splitting the control and data planes between macro base stations and small cells respectively, such that macro base stations handle control signals, while small cells handle data [108]. Other technologies such as Wi-Fi offloading can also be integrated into HetNets.

Appendix A

Proofs

A.1 Reducing Generalized Inverse Calculation Complexity (Section 2.2.3.1)

Computing $\mathbf{W} = \mathbf{G} + \mathbf{UB}$ using $\mathbf{B} = -(\mathbf{h}^H\mathbf{U})^{-1}\mathbf{h}^H\mathbf{G}$ involves a psuedoinverse, inverse and nulllspace calculation. We propose a reduced complexity computation method to calculate the same $\mathbf{W}$:

Claim: The first $N-1$ columns of $\begin{pmatrix}\mathbf{H}^H\\ \mathbf{h}^H\end{pmatrix}^+$ is equal to $\mathbf{W} = \mathbf{G} + \mathbf{UB}$ if $\mathbf{B} = -(\mathbf{h}^H\mathbf{U})^{-1}\mathbf{h}^H\mathbf{G}$. That is, $\begin{pmatrix}\mathbf{H}^H\\ \mathbf{h}^H\end{pmatrix}^+ = \begin{pmatrix}\mathbf{W} & \mathbf{z}\end{pmatrix}$ where $\mathbf{z}$ is a particular column vector.

Proof: According to [109], for a partitioned matrix $\mathbf{A} = \begin{pmatrix}\mathbf{A_1} & \mathbf{A_2}\end{pmatrix}$ of dimension $m \times n$,

$$\mathbf{A}^+ = \begin{pmatrix}\mathbf{A_1}^+ - \mathbf{A_1}^+\mathbf{A_2}\left((\mathbf{I}_m - \mathbf{A_1}\mathbf{A_1}^+)\mathbf{A_2}\right)^+\\ \mathbf{A_2}^+ - \mathbf{A_2}^+\mathbf{A_1}\left((\mathbf{I}_m - \mathbf{A_2}\mathbf{A_2}^+)\mathbf{A_1}\right)^+\end{pmatrix} \tag{A.1}$$

Therefore, by extension, for a matrix $\mathbf{A}^H = \begin{pmatrix}\mathbf{A_1}^H\\ \mathbf{A_2}^H\end{pmatrix}$ of dimension $m \times n$,

$$(\mathbf{A}^H)^+ = \begin{pmatrix}(\mathbf{A_1}^H)^+ - \left(\mathbf{A_2}^H(\mathbf{I}_m - \mathbf{A_1}\mathbf{A_1}^+)^H\right)^+\mathbf{A_2}^H(\mathbf{A_1}^H)^+,\\ (\mathbf{A_2}^H)^+ - \left(\mathbf{A_1}^H(\mathbf{I}_m - \mathbf{A_2}\mathbf{A_2}^+)^H\right)^+\mathbf{A_1}^H(\mathbf{A_2}^H)^+\end{pmatrix} \tag{A.2}$$

Now, expanding $\mathbf{W}$,

$$\mathbf{W} = \mathbf{G} - \mathbf{U}(\mathbf{h}^H\mathbf{U})^{-1}\mathbf{h}^H\mathbf{G} = \left(\mathbf{I}_N - \mathbf{U}(\mathbf{h}^H\mathbf{U})^{-1}\mathbf{h}^H\right)\mathbf{G}. \tag{A.3}$$

Substituting $\mathbf{A_1} = \mathbf{H}$ and $\mathbf{A_2} = \mathbf{h}$ into (8), we see that the left matrix becomes

$$\mathbf{G} - \left(\mathbf{h}^H(\mathbf{I}_m - \mathbf{H}\mathbf{H}^+)^H\right)^+\mathbf{h}^H\mathbf{G} = \left(\mathbf{I}_N - \left(\mathbf{h}^H(\mathbf{I}_N - \mathbf{H}\mathbf{H}^+)^H\right)^+\mathbf{h}^H\right)\mathbf{G}. \tag{A.4}$$

Since $\mathbf{I}_N - \mathbf{H}\mathbf{H}^+ = \mathbf{U}\mathbf{U}^H$,

$$\begin{aligned}\left(\mathbf{I}_N - \left(\mathbf{h}^H(\mathbf{I}_N - \mathbf{H}\mathbf{H}^+)^H\right)^+\mathbf{h}^H\right)\mathbf{G} &= \left(\mathbf{I}_N - \left(\mathbf{h}^H(\mathbf{U}\mathbf{U}^H)^H\right)^+\mathbf{h}^H\right)\mathbf{G} \\ &= \left(\mathbf{I}_N - \left(\mathbf{h}^H(\mathbf{U}\mathbf{U}^H)\right)^+\mathbf{h}^H\right)\mathbf{G}.\end{aligned} \tag{A.5}$$

Notice that

$$\left(\mathbf{h}^H(\mathbf{U}\mathbf{U}^H)\right)^+ = \left((\mathbf{h}^H\mathbf{U})\mathbf{U}^H\right)^+ = (\mathbf{U}^H)^+(\mathbf{h}^H\mathbf{U})^{-1} = \mathbf{U}(\mathbf{h}^H\mathbf{U})^{-1} \tag{A.6}$$

using the properties of the psuedoinverse, namely:

- For any full column and row rank matricies $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{A}\mathbf{B})^+ = \mathbf{B}^+\mathbf{A}^+$.
- For an invertible square matrix, $(\cdot)^+ = (\cdot)^{-1}$.
- If a matrix $\mathbf{A}$ of dimension $m \times n$ has orthonormal columns, i.e.$\mathbf{A}^H\mathbf{A} = \mathbf{I}_n$, then $\mathbf{A}^+ = \mathbf{A}^H$.

Substituting (12) into (11) will give (9), and thus proving our claim.

A.2 Effect of Imperfect CSI on Generalized Inverse Precoder (Proposition 2.1)

Suppose the last row of a matrix is modified or corrupted, and its inverse is calculated. We investigate the effects of using this corrupted inverse on the original matrix. In the context of wireless communications, using a ZF precoder constructed using imperfect CSI (each channel defined as rows in system equations) will only affect the performance of that imperfect user. Other users do not suffer any degra-

dation.

Claim: Multiplying the corrupted inverse with the original matrix will result in an identity matrix except for the corrupted row.

Proof: Consider a square invertible matrix $\begin{pmatrix} \mathbf{H} \\ \mathbf{h}+\mathbf{\Delta h} \end{pmatrix}$ where $\mathbf{h}$ is a row vector and $\mathbf{\Delta h}$ is the error. Using the matrix identity $(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}(\mathbf{I}+\mathbf{B}\mathbf{A}^{-1})^{-1}$, we have

$$\begin{pmatrix} \mathbf{H} \\ \mathbf{h}+\mathbf{\Delta h} \end{pmatrix}^{-1} = \left(\begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \end{pmatrix} \right)^{-1} \tag{A.7}$$

$$= \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} - \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \end{pmatrix} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \left(\mathbf{I} + \begin{pmatrix} \mathbf{0} \\ \Delta \mathbf{h} \end{pmatrix} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \right)^{-1}. \tag{A.8}$$

Thus, right multiplying this corrupted inverse with the correct original matrix,

$$\begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix} \begin{pmatrix} \mathbf{H} \\ \mathbf{h}+\mathbf{\Delta h} \end{pmatrix}^{-1} = \mathbf{I} - \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \end{pmatrix} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \left(\mathbf{I} + \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \end{pmatrix} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \right)^{-1}$$

$$= \mathbf{I} - \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \left(\mathbf{I} + \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \end{pmatrix} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \right)^{-1} \end{pmatrix} \tag{A.9}$$

$$= \mathbf{I} - \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \left(\mathbf{I} + \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta h} \begin{pmatrix} \mathbf{H} \\ \mathbf{h} \end{pmatrix}^{-1} \end{pmatrix} \right)^{-1} \end{pmatrix}. \tag{A.10}$$

This indicates that only the bottom corrupted row will become a non-identity row, while the other rows will remain that of identity matrix form. This proof can be applied without loss of generality to any particular row.

In general, if $\mathbf{H}$ also contains an error,

$$\begin{pmatrix}\mathbf{H}\\ \mathbf{h}\end{pmatrix}\begin{pmatrix}\mathbf{H}+\mathbf{\Delta H}\\ \mathbf{h}+\mathbf{\Delta h}\end{pmatrix}^{-1} = \mathbf{I} - \begin{pmatrix}\mathbf{\Delta H}\begin{pmatrix}\mathbf{H}\\ \mathbf{h}\end{pmatrix}^{-1}\mathbf{Z}\\ \mathbf{\Delta h}\begin{pmatrix}\mathbf{H}\\ \mathbf{h}\end{pmatrix}^{-1}\mathbf{Z}\end{pmatrix} \tag{A.11}$$

where

$$\mathbf{Z} = \left(\mathbf{I} + \begin{pmatrix}\mathbf{\Delta H}\begin{pmatrix}\mathbf{H}\\ \mathbf{h}\end{pmatrix}^{-1}\\ \mathbf{\Delta h}\begin{pmatrix}\mathbf{H}\\ \mathbf{h}\end{pmatrix}^{-1}\end{pmatrix}\right)^{-1}, \tag{A.12}$$

indicating that errors in both $\mathbf{H}$ and $\mathbf{h}$ will affect all rows.

A.3 Tikhonov Regularization Parameter and Constraint Relationship (Proposition 2.2)

The general form of a least squares problem with a quadratic constraint and its Tikhonov equivalent is

$$\min \|Ax-b\|^2 \text{ s.t. } \|Cx-d\|^2 \leq \alpha^2 \rightarrow \min \|Ax-b\|^2 + \lambda^2 \|Cx-d\|^2 \tag{A.13}$$

with the solution

$$x = (A^H A + \lambda^2 C^H C)^{-1}(A^H b + C^H d). \tag{A.14}$$

Conventional definitions state that A and C are matrices, x, b and d are vectors, but we can extend the solution without loss of generality to suitably dimensioned vectors or matrices for any of the variables.

We can set $A = \widetilde{\mathbf{H}}^H\mathbf{U}$, $b = -\widetilde{\mathbf{H}}^H\mathbf{G}$, $C = \mathbf{U}$, $d = -\mathbf{G}$, $x = \mathbf{B}$ to obtain the objective

$$\min \left\|\widetilde{\mathbf{H}}^H(\mathbf{G}+\mathbf{UB})\right\|^2 + \lambda^2 \|\mathbf{G}+\mathbf{UB}\|^2. \tag{A.15}$$

The relationship between the constraints α, β and parameter λ relies on the generalized singular value decomposition (GSVD) of $\widetilde{\mathbf{H}}^H\mathbf{U}$ and $\mathbf{U}$. To show this, consider the expression for the solution to (A.15) using (A.14) [60], which is

$$\mathbf{B} = \left((\widetilde{\mathbf{H}}^H\mathbf{U})^H\widetilde{\mathbf{H}}^H\mathbf{U} + \lambda^2\mathbf{U}^H\mathbf{U}\right)^{-1}\left((\widetilde{\mathbf{H}}^H\mathbf{U})^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + \mathbf{U}^H(-\mathbf{G})\right).$$

To first prove the relationship between α and λ, let the GSVD of $\widetilde{\mathbf{H}}^H\mathbf{U}$ and $\mathbf{U}$ be

$$\widetilde{\mathbf{H}}^H\mathbf{U} = \mathbf{L}_1\mathbf{\Sigma}\mathbf{R}^{-1} \text{ and } \mathbf{U} = \mathbf{L}_2\mathbf{M}\mathbf{R}^{-1}, \tag{A.16}$$

where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$ and $M = \text{diag}(\mu_1, \ldots, \mu_{N-k})$. Note that $\mathbf{L}_1$ and $\mathbf{L}_2$ are unitary matrices, i.e., $\mathbf{L}_1^H\mathbf{L}_1 = \mathbf{L}_2^H\mathbf{L}_2 = \mathbf{I}$.

Substituting the above into (A.16) and simplifying, we have

$$\begin{aligned}
\mathbf{B} =& \left((\mathbf{L}_1\mathbf{\Sigma}\mathbf{R}^{-1})^H\mathbf{L}_1\mathbf{\Sigma}\mathbf{R}^{-1} + \lambda^2(\mathbf{L}_2\mathbf{M}\mathbf{R}^{-1})^H\mathbf{L}_2\mathbf{M}\mathbf{R}^{-1}\right)^{-1} \\
&\times \left((\mathbf{L}_1\mathbf{\Sigma}\mathbf{R}^{-1})^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + (\mathbf{L}_2\mathbf{M}\mathbf{R}^{-1})^H(-\mathbf{G})\right) \\
=& \left((\mathbf{R}^{-1})^H\mathbf{\Sigma}^H\mathbf{L}_1^H\mathbf{L}_1\mathbf{\Sigma}\mathbf{R}^{-1} + \lambda^2(\mathbf{R}^{-1})^H\mathbf{M}^H\mathbf{L}_2^H\mathbf{L}_2\mathbf{M}\mathbf{R}^{-1}\right)^{-1} \\
&\times \left((\mathbf{R}^{-1})^H\mathbf{\Sigma}^H\mathbf{L}_1^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + (\mathbf{R}^{-1})^H\mathbf{M}^H\mathbf{L}_2^H(-\mathbf{G})\right) \\
=& \left((\mathbf{R}^{-1})^H(\mathbf{\Sigma}^H\mathbf{L}_1^H\mathbf{L}_1\mathbf{\Sigma} + \lambda^2\mathbf{M}^H\mathbf{L}_2^H\mathbf{L}_2\mathbf{M})\mathbf{R}^{-1}\right)^{-1} \\
&\times (\mathbf{R}^{-1})^H\left(\mathbf{\Sigma}^H\mathbf{L}_1^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + \mathbf{M}^H\mathbf{L}_2^H(-\mathbf{G})\right) \\
=& \mathbf{R}(\mathbf{\Sigma}^H\mathbf{\Sigma} + \lambda^2\mathbf{M}^H\mathbf{M})^{-1}\left(\mathbf{\Sigma}^H\mathbf{L}_1^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + \mathbf{M}^H\mathbf{L}_2^H(-\mathbf{G})\right).
\end{aligned} \tag{A.17}$$

Setting the equality constraint in (2.20) to $\|Cx - d\| = \alpha$ and substituting the above, we have

$$\begin{aligned}
\alpha &= \left\|\mathbf{L}_2\mathbf{M}^{-1}(\mathbf{\Sigma}^H\mathbf{\Sigma} + \lambda^2\mathbf{M}^H\mathbf{M})^{-1}\left(\mathbf{\Sigma}^H\mathbf{L}_1^H(-\widetilde{\mathbf{H}}^H\mathbf{G}) + \mathbf{M}^H\mathbf{L}_2^H(-\mathbf{G})\right) - \mathbf{L}_2^H(-\mathbf{G})\right\| \\
&= \left\|\frac{\mathbf{\Omega}}{\sum_i \sigma_i^2 + \lambda^2\sum_i \mu_i^2} - \mathbf{\Psi}\right\|.
\end{aligned} \tag{A.18}$$

Note that we obtained the denominator in the above expression by using the fact that $\mathbf{\Sigma}^H\mathbf{\Sigma}$ and $\mathbf{M}^H\mathbf{M}$ are diagonal matrices, and thus can be easily simplified. We have

also utilised the fact that $\mathbf{L}_2$ is a unitary matrix.

The above equation is difficult to manipulate as $\mathbf{\Omega}$ and $\mathbf{\Psi}$ are matrices. However, since a vectorized matrix has the same norm as the original matrix as it contains the same elements, we can simplify the above further to obtain (A.20):

$$\alpha = \left\| \frac{\text{vec}(\mathbf{\Omega})}{\sum_i \sigma_i^2 + \lambda^2 \sum_i \mu_i^2} - \text{vec}(\mathbf{\Psi}) \right\| = \frac{\sum_i \omega_i}{\sum_i \sigma_i^2 + \lambda^2 \sum_i \mu_i^2} - \sum_i \psi_i. \tag{A.19}$$

We can rearrange to obtain

$$\lambda^2 = \frac{1}{\sum_i \mu_i^2} \left(\frac{\sum_i \omega_i}{\alpha + \sum_i \psi_i} - \sum_i \sigma_i^2 \right). \tag{A.20}$$

A.4 Quasiconvexity of Sum SINR (Proposition 6.1)

From [110], a differentiable function is quasiconvex if and only if

$$f(y_1, \ldots, y_n) \leq f(x_1, \ldots, x_n) \Rightarrow \nabla f(x_1, \ldots, x_n)^T (y_1 - x_1, \ldots, y_n - x_n)^T \leq 0. \tag{A.21}$$

Although we can apply this to the sum SINR function directly, we note that since the addition of bounds and differentiation are preserved under addition, it is sufficient to show that each SINR is quasiconvex in order to show that sum SINR is quasiconvex[9]. Further, we can ignore the noise constant in the denominator as it does not change the convexity behaviour or shape of each fraction.

Consider the generic definition of sum SINR

$$S = \sum_{i=1}^{N} \frac{P_i}{a_i} \tag{A.22}$$

where $a_i = \sum_{j \neq i} P_j + \sigma^2$. For N varying powers, if the numerator of an SINR fraction is constant, i.e., a power that is not varying, then

$$\frac{k}{P_1 + \ldots + P_N} \tag{A.23}$$

[9] In general, the sum of quasi-convex functions may not be quasi-convex.

for varying powers $P_1, \ldots, P_N$ and constant k is clearly quasiconvex as it follows a hyperbolic shape. If the numerator is a varying power, i.e.,

$$\frac{P_1}{P_2 + \ldots + P_{N-1}}, \tag{A.24}$$

then using the second inequality in (A.21) we find that for two sets of powers $\{P_1, \ldots, P_N\}$ and $\{P_1', \ldots, P_N'\}$,

$$\nabla S(P_1, \ldots, P_n)^T (P_1' - P_1, \ldots, P_n' - P_n)^T = \begin{pmatrix} \frac{1}{P_2 + \ldots + P_N} & \cdots & \frac{-P_1}{(P_2 + \ldots + P_N)^2} \end{pmatrix} \begin{pmatrix} P_1' - P_1 \\ \vdots \\ P_N' - P_N \end{pmatrix} =$$

$$\frac{P_1' - P_1}{P_2 + \ldots + P_N} - P_1 \sum_{i=2}^{N} \frac{P_i' - P_i}{(P_2 + \ldots + P_N)^2} \leq 0,$$

$$(P_1' - P_1)(P_2 + \ldots + P_N) - P_1 \sum_{i=2}^{N} (P_i' - P_i) \leq 0,$$

$$P_1'(P_2 + \ldots + P_N) \leq P_1(P_2' + \ldots + P_N'),$$

$$\frac{P_1'}{P_2' + \ldots + P_N'} \leq \frac{P_1}{P_2 + \ldots + P_N}. \tag{A.25}$$

This is the first inequality in (A.21) when $P_i' = y_i$, $P_i = x_i$. Thus, for any combination of varying powers, we find that sum SINR is quasiconvex.

A.5 Maximizing Sum Rate and Sum SINR (Proposition 6.2)

Suppose we differentiate S in (A.22) with respect to the most dominant power P_i:

$$\frac{dS}{dP_i} = \frac{1}{a_i} - \sum_{j \neq i}^{N} \frac{P_j}{a_j^2}. \tag{A.26}$$

If P_i was a dominant power, we observe that the derivative will approach $1/a_i$ since $a_j \to \infty$ as $P_i \to \infty$.

Similarly, for the generic definition of sum rate

$$\mathcal{R} = \sum_{i=1}^{N} \log_2 \left(1 + \frac{P_i}{\sum_{j \neq i} P_j + \sigma^2}\right) = \log_2 \prod_{i=1}^{N} \left(1 + \frac{P_i}{\sum_{j \neq i} P_j + \sigma^2}\right), \tag{A.27}$$

if we expand out the brackets in (A.27) ignoring the logarithm to obtain

$$\begin{aligned} \prod_{i=1}^{N} \left(1 + \frac{P_i}{a_i}\right) &= 1 + \sum_{i=1}^{N} \frac{P_i}{a_i} + \sum(\text{Products of SINRs two at a time}) \\ &+ \sum(\text{Products of SINRs three at a time}) + \ldots + \prod_{i=1}^{N} \frac{P_i}{a_i} \end{aligned} \tag{A.28}$$

and differentiate with respect to P_i, we find that all the derivatives of the products of SINRs will contain a_j^2 in the denominator, and will approach 0 as $P_i \to \infty$. Thus, both sum SINR and (A.28) have the same *asymptotic gradient* of $1/a_i$ when one power dominates. Note that if we differentiate with respect to P_i, but P_i was not the dominate power, both expressions will instead approach $-P_j/a_j^2$ if P_j was the dominant power.

A.6 General Solution for Unconstrained Frequency Sharing in Dedicated Mode (Section 6.4.2.1)

For N transmitters, and hence N partitions, sum rate is

$$\mathcal{R} = \sum_{i=1}^{N} \alpha_i \log_2 \left(1 + \frac{\gamma_i}{\alpha_i}\right), \tag{A.29}$$

where $\sum \alpha_i = 1$ is the partition fraction and γ_i is the SNR of each receiver.

In order to greedily maximize R, we need to simultaneously solve the partial derivatives with respect to each α_i, i.e. $\frac{\partial \mathcal{R}}{\partial \alpha_i} = 0$. This will give the relations

$$\frac{\gamma_i}{\alpha_i} = \frac{\gamma_k}{\alpha_k} \tag{A.30}$$

for $i, k = 1, \ldots, N$. Setting $k = m$, and noting that $\alpha_m = 1 - \sum_{k=1}^{N-1} \alpha_k$, we can rearrange

(A.30) to obtain

$$\alpha_i = \left(1 - \sum_{k=1}^{N-1} \frac{\alpha_i \gamma_k}{\gamma_i}\right) \frac{\gamma_i}{\gamma_n} = \frac{\gamma_i}{\gamma_m} - \frac{\alpha_i}{\gamma_m} \sum_{k=1}^{N-1} \gamma_k, \tag{A.31}$$

which can be simplified to

$$\alpha_i = \frac{\gamma_i}{\gamma_m + \sum_{k=1}^{N-1} \gamma_k} = \frac{\gamma_i}{\sum_{k=1}^{N} \gamma_k}. \tag{A.32}$$

Thus, each resource partition fraction is equal to the fraction of the particular SNR over the total SNR. Substituting the above into (A.29) will always give the maximum sum rate

$$\mathcal{R} = \log_2 \left(1 + \sum_{i=1}^{N} \gamma_i\right). \tag{A.33}$$

A.7 Closed Form Solution for Constrained Frequency Sharing in Dedicated D2D Mode (Section 6.4.2.2)

To solve (6.19), we need to manipulate (6.19) to a form where we can use the Lambert W function. Firstly, we can rearrange and then exponentiate (6.19) to get

$$\ln\left(1 + \frac{\gamma_r}{\alpha}\right) = e^{\mathcal{R}_r^{\min} \ln 2/\alpha}. \tag{A.34}$$

Next, we need to introduce additional terms such that the exponent contains the left hand side, i.e.,

$$\ln\left(1 + \frac{\gamma_r}{\alpha}\right) = 2^{-\mathcal{R}_r^{\min}/\gamma_r} e^{\frac{\mathcal{R}_r^{\min} \ln 2}{\gamma_r}\left(1 + \frac{\gamma_r}{\alpha}\right)}. \tag{A.35}$$

Moving the exponential to the left hand side gives

$$-\frac{\mathcal{R}_r^{\min} \ln 2}{\gamma_r} \ln\left(1 + \frac{\gamma_r}{\alpha}\right) e^{-\frac{\mathcal{R}_r^{\min} \ln 2}{\gamma_r}\left(1 + \frac{\gamma_r}{\alpha}\right)} = -\frac{\mathcal{R}_r^{\min} \ln 2}{\gamma_r} 2^{-\mathcal{R}_r^{\min}/\gamma_r}. \tag{A.36}$$

We can now apply the Lambert W function since the exponential is in the form Ae^A:

$$-\frac{\mathcal{R}_r^{\min} \ln 2}{\gamma_r} \ln\left(1 + \frac{\gamma_r}{\alpha}\right) = \mathrm{W}\left(-\frac{\mathcal{R}_r^{\min} \ln 2}{\gamma_r} 2^{-\mathcal{R}_r^{\min}/\gamma_r}\right). \tag{A.37}$$

Rearranging for α gives the solution

$$\alpha = \frac{-\gamma_r \mathcal{R}_r^{\min} \ln 2}{\mathcal{R}_r^{\min} \ln 2 + \gamma_r \mathrm{W}\left(-\frac{\mathcal{R}_r^{\min} \ln 2}{\gamma_r} 2^{-\mathcal{R}_r^{\min}/\gamma_r}\right)}. \tag{A.38}$$

To ensure a real solution, we use the -1 branch of the Lambert W function.

Appendix B

Geometric Solution for Power Control

Key Question: *Will the same set of powers that maximize sum SINR also maximize sum rate?*

This appendix generalizes the power reuse problem in D2D Chapter 6 to N users and transmitters, and illustrates the accuracy and closeness of using sum SINR as an approximation to sum rate when one received power is dominant over others.

The general multi-user interference channel allows multiple transmitters and receivers to communicate simultaneously, but requires power control to ensure QoS for each user, while at the same time maximizing some system metric. While capacity is the ideal metric from an information theory perspective, the capacity of a multiuser interference channel is still unknown [111]. Alternatively, researchers have focused on sum rate as a metric.

Although it has been proven for two transmitting sources that sum rate is a convex expression [112], it is known that for more than two sources sum rate is generally non-convex, and thus not easy to maximize using standard optimization techniques. Often, non-convex formulations can be transformed into more manageable forms, as is the case with geometric programming [88, 103], but these require certain approximations (e.g., high SINR regime).

An early approach to the power control problem was through finding a Pareto optimal solution [113]. However, this does not necessarily maximize the objective, nor does it always satisfy individual power constraints. Such an approach has been used as an initial feasibility test [114].

An alternative for sum rate for more than two users is to use sum SINR [115], which is easier to maximize due to the absence of a log term. Although for one or two users this will lead to the same optimal solutions, in general, maximizing sum SINR may not also maximize sum rate. It is known that sum rate is not convex with respect to arbitrary combinations of varying powers (i.e., not jointly convex in all powers), but neither is sum SINR, which is known to be convex with one varying power but not convex in general [116]. There has been no discussion or mathematical study on the relationship between sum rate and sum SINR and their behaviour with varying powers.

In [88], the authors proved that to maximize sum rate with individual power constraints, binary power control, i.e., each power operates either at maximum or minimum levels, is the optimal solution for two users, and a suboptimal solution for more than two users. Further, the authors indirectly suggest that binary power control is the optimal solution to any objective that is convex. A bound on the approximation of sum rate with an alternative expression relying on the arithmetic-geometric mean inequality and the Specht's ratio is also given, but this does not answer the question of whether the same set of powers can maximize both sum rate and sum SINR, or how similar are the powers that do. Further, [88] does not include individual rate constraints.

We describe graphically the feasible power region under both individual power and rate (or equivalently, SINR) constraints, and show the accuracy of using the vertices or corners of this region as solutions to maximize sum rate.

B.1 System Model

Consider a system with N links, each of which has a unique transmitter and receiver. Each receiver treats any interference it receives from the other links as noise. We

desire to solve:

$$\text{maximize} \quad R = \sum_{i=1}^{N} \log_2 \left(1 + \frac{h_{i,i}p_i}{\sum_{j\neq i} h_{j,i}p_j + \sigma^2}\right) = \log_2 \left(\prod_{i=1}^{N} \left(1 + \frac{h_{i,i}p_i}{\sum_{j\neq i} h_{j,i}p_j + \sigma^2}\right)\right) \tag{B.1}$$

$$\text{subject to} \quad p_i \leq P_i^{\max}, \tag{B.2}$$

$$\frac{h_{i,i}p_i}{\sum_{j\neq i} h_{j,i}p_j + \sigma^2} \geq \gamma_i, \; i = 1, \ldots, N. \tag{B.3}$$

where $h_{j,i}$ is the channel gain from the jth transmitter to the ith receiver, p_i is the transmission power at the ith transmitter, $P_i^{\max}$ is the maximum transmission power at the ith transmitter, γ_i is the SINR threshold corresponding to the minimum rate for the ith user, and σ^2 is the zero mean additive Gaussian white noise (AWGN). Equations (B.2) and (B.3) represent the individual transmit power and minimum user rate constraints respectively.

For analytical simplicity, we drop the channel gains $h_{j,i}$, as fading characteristics become less significant compared to differences in magnitudes of transmit powers. Thus, p_i represent the received powers from each transmitter. For example, a macro station may transmit at 43 dB compared to 23 dB for a femtocell, where the 20 dB difference in magnitude will mostly dominate fading effects. Our conclusions are therefore more accurate for Gaussian channels. We also let $a_i = \sum_{j\neq i} p_j + \sigma^2$ to represent the interference plus noise at the ith receiver.

B.2 Power Region in N-dimensions

Plotting individual power constraints on their own orthogonal axis in N-dimensional space $\mathbb{R}^N$, the feasible power region can be described as a hypercube, the interior and boundary of which contains all possible transmit powers. The corners or vertices of the hypercube are the points with coordinates $(p_1, \ldots, p_N)$ either $p_i = 0$ or $p_i = P_i^{\max}$. For different maximum powers, e.g., in a downlink heterogeneous network, the hypercube will have different side lengths.

In addition to individual power constraints, the feasible region can be formed

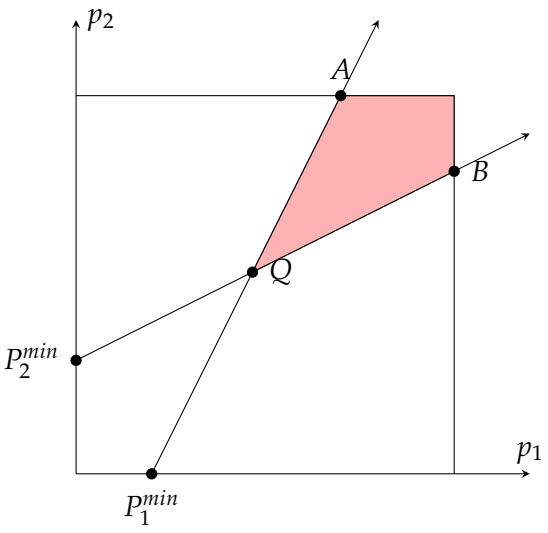

Figure B.1: Power region for two transmitters bound by edges of the rectangle (power constraint) and lines (minimum rate constraint).

by minimum user rate constraints. By rearranging the minimum rate constraints in (B.3), we can obtain N inequalities of the form

$$p_i - \gamma_i \left(\sum_{j \neq i} p_j \right) \geq \gamma_i \sigma^2, \forall i \in 1, \ldots, N. \tag{B.4}$$

Geometrically, with equality the above is the equation of a hyperplane in $\mathbb{R}^N$, while with inequality it is the region above[10] the hyperplane. Thus, *the power constraints form the hypercube, while the minimum user rate constraints further bound the power region into a polytope*. Increasing the number of powers increases the dimensionality of the region, while increasing the number of users increases the number of hyperplanes and further restricts the polytope.

The intersection of all the SINR inequalities, denoted as the point Q, can be found by solving for their equality expressions simultaneously, which can be done using methods such as Cramer's Rule. The final region bounded by the boundaries of the hypercube and the hyperplanes form the feasible power region. Possible regions for two and three transmitting powers are illustrated in Figs. B.1 and B.2.

[10]Here, 'above' refers to the region satisfying the inequality, and may not always be 'above' in the everyday sense.

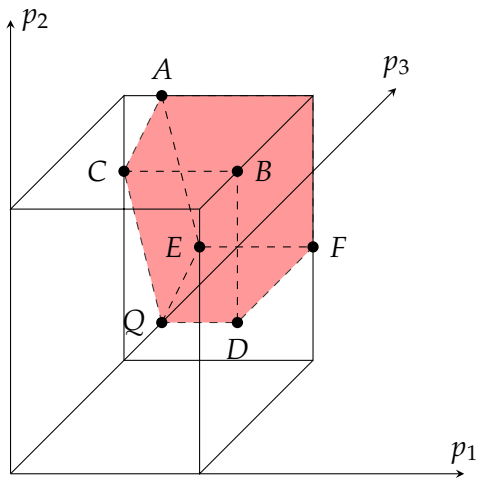

Figure B.2: Power region for three transmitters bound by edges of the cube (power constraint) and planes (minimum rate constraint, not shown for clarity).

B.3 Sum SINR Approximation

From [88], we know that the optimal powers that maximizes sum rate occur at the boundary of the power region. Further, if the function to be maximized is convex, the vertices of the power region, with the exception of the point Q, form the finite set of points that contain the optimal powers. The coordinates of these vertices are powers that are either maximum powers defined by the power constraints, or minimum powers allowed by other users' constraints. However, although convexity is a more common property to prove, it is in fact *quasiconvexity* which states that for a given domain, the maximum of a function lies on the endpoints. Of course, convex functions are also quasiconvex and share this property.

Previously, other works have claimed that since SINRs are convex expressions in individual powers, which implies that the optimal powers lie on the vertices. However, this only applies if the vertices also lie on the hypercube edges of the power region, since on the edges only one power is varying. For N powers, the hypercube will have vertices that lie on other types of boundaries, e.g. a face, which represent the condition that there is more than one varying power. Hence, on these boundaries where more than one power is varying, it is also required that the function is also quasiconvex with respect to more than one varying power in order to justify

searching only vertices for optimal powers.

To illustrate, consider the simple case of a two user system in Fig. B.1. Without minimum user constraints, binary power control tells us the powers which will maximize sum rate will either be $p_i = P_i^{\max}$ or $p_i = 0$. However, with the minimum user constraints, the power region is now also bounded by the lines, and thus the optimal set of powers also include the points A and B, whose coordinates can be found by using (B.4). A similar approach can be used to determine the set of points for $N \geq 4$ dimensions, although they become increasingly more difficult to visualize.

From Chapter 6 and Section A.4, we have established that sum SINR can be a close approximation of sum rate when one received power is an order of magnitude larger than the others (Proposition 6.2) due to its quasiconvexity and derivative behaviour. We can further show that although sum rate is not convex in general, it is convex with respect to one power, i.e.,

Proposition B.1. *For any number of transmitting powers, sum rate in individual powers, i.e., one power varying and the others constant, is always convex.*

Proof. Consider the expression within the $\log_2$ in (A.27), i.e.,

$$\prod_{i=1}^{N}\left(1+\frac{p_i}{a_i}\right)=\frac{(\sum_{i=1}^{N}p_i+\sigma^2)^N}{\prod_{i=1}^{N}a_i}=\frac{(x+a_i)^N}{a_i\prod_{k\neq i}^{N-1}(x+a_{i,k})}\triangleq f(x) \tag{B.5}$$

where $x = p_i$, $a_i = \sum_{j\neq i} p_j + \sigma^2$ and $a_{i,k} = \sum_{j\neq i,k} p_j + \sigma^2$. The N roots of $f(x)$ are at $x = -a_i$, while the asymptotes are at $x = -a_{i,k}$ for $k \neq i$. Since $a_i = a_{i,k} + p_k$, $a_i > a_{i,k}$, meaning that the roots occur to the left of all the asymptotes. To show that $f(x)$ is convex for $x > 0$, we can take derivatives and use the precise definition of convexity, but this is tedious to do with so many products. Instead, we adopt a graphical approach.

In general, since $f(x)$ is a function with polynomial numerators and denominators, basic curve sketching techniques can be employed to determine its generic shape.

1. Consider the case when N is even (Fig. B.3). The smallest, i.e., left most critical point is the root at $x = -a_i$. If N is even $f(x)$ must have either a maximum

or minimum turning point at that root. It is easy to see that since for x to the left of the first vertical asymptote, $f(x) \leq 0$, the function must have a maximum turning point at $x = -a_i$. The behaviour and shape of $f(x)$ then alternates between convex positive and concave negative graphs between each set of asymptotes. Since there are an even number of such graphs, the right most one corresponding to when $x > 0$ will always be positive and convex.

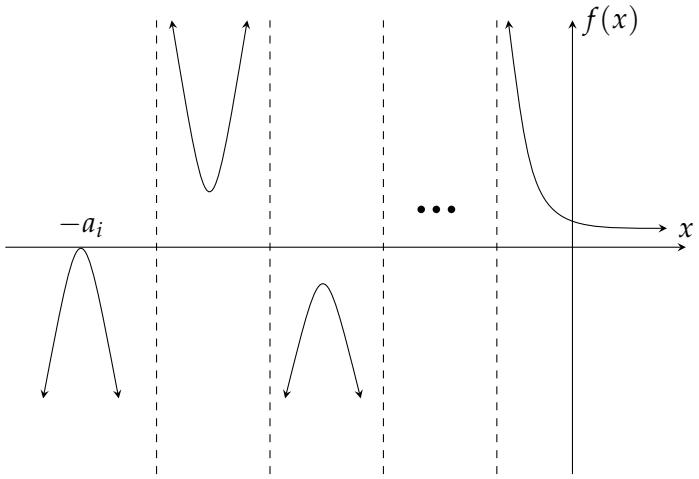

Figure B.3: General curve behaviour of sum rate with respect to one power for even number of powers.

2. Consider the case when N is odd (Fig. B.4). At the root, the function has a point of inflexion due to the odd power, while it is easy to see that $f(x)$ will be negative between when x is between the two left-most vertical asymptotes. Following the same pattern as the even case, the function will alternate between convex positive and concave negative graphs between each set of asymptote, and again will end up being positively convex for $x > 0$.

Since $\log_2$ is a monotonically increasing function, and the relevant branches are decreasing with second derivatives less than 0, the sum rate over those ranges will remain convex. ■ ■

Remark B.1. The convexity of sum rate with one varying power means that known convex methods can be used to solve for power along an edge of the power region.

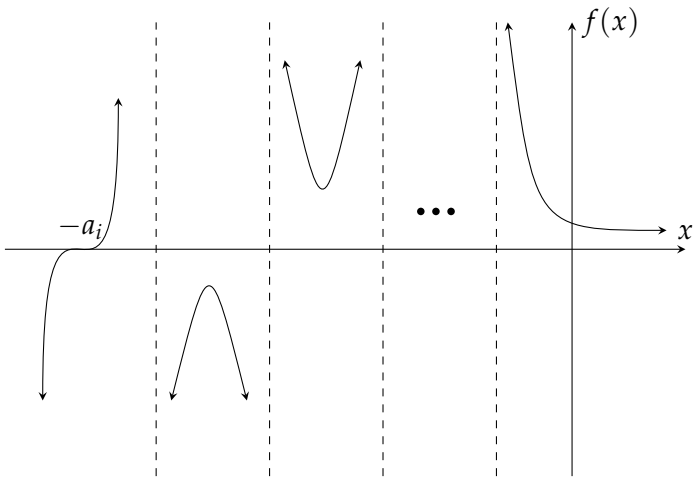

Figure B.4: General curve behaviour of sum rate with respect to one power for odd number of powers.

B.4 Simulation Results

In our simulations we considered three and four transmitting powers, and tested all combinations to illustrate the validity of Proposition 6.2. All powers were normalized with respect to the noise power. We set a power range for each transmitter, and tested all combinations of powers with step sizes chosen such that there were five powers in each transmitting set. All possible combinations of powers were searched through, with each combination labelled with a 'search index.'

For three transmitters, Fig. B.5 shows the derivatives of sum SINR and the product term in (A.27) when received powers are of the same order of magnitude around 10 dB with respect to the noise power, while Fig. B.6 shows the derivatives when one power is an order of magnitude larger than others. It is clear that when there is one dominating power, the derivatives coincide almost perfectly, indicating that the log term in sum rate and sum SINR, 'follow' one another and thus have their maxima and minima occur at the same locations. We observe the same trend when there are four transmitters as shown in Fig. B.7 and Fig. B.8.

When considering the actual sum rate, i.e., taking the logarithm, we find that the global maxima and minima indeed still occur at the same set of powers as expected when a received power is an order of magnitude larger, as shown in Fig. B.9 for three

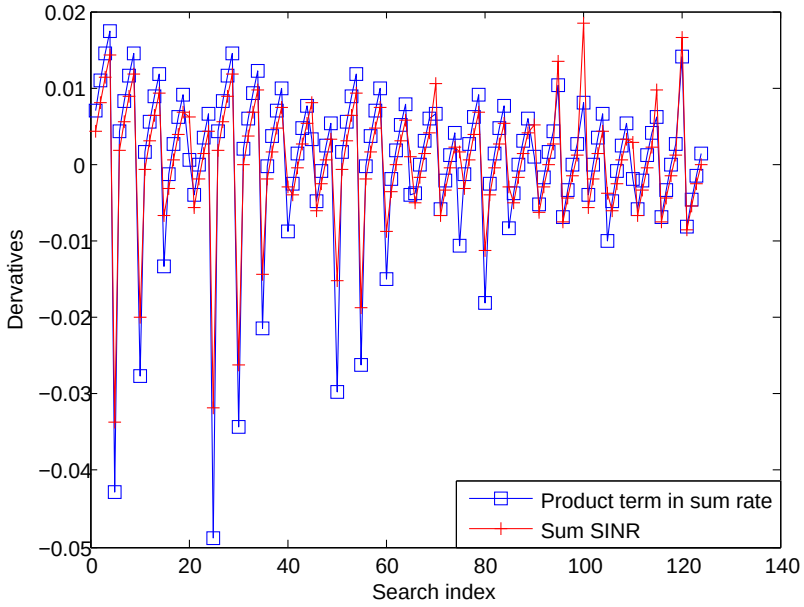

Figure B.5: Powers the same order of magnitude. There is a mismatch of derivative values with no consistency.

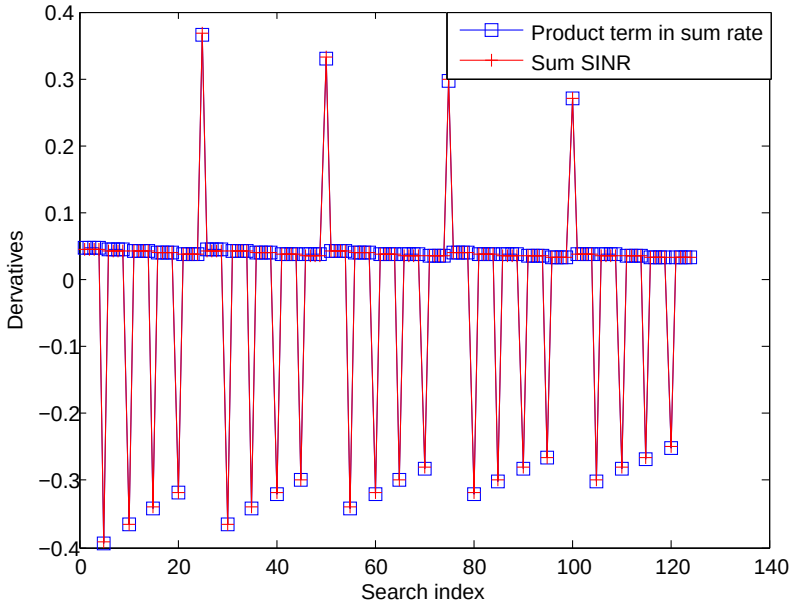

Figure B.6: One power an order of magnitude larger. Derivative values match almost perfectly.

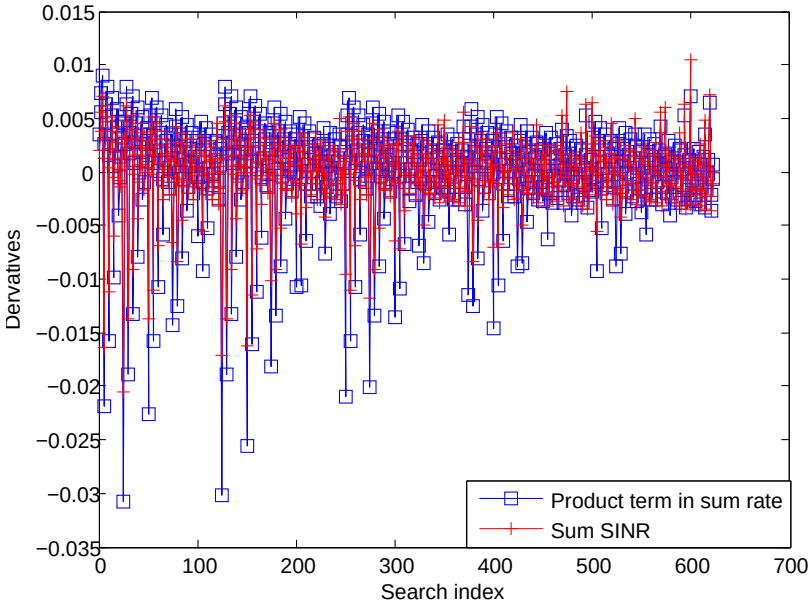

Figure B.7: Powers the same order of magnitude. There is a mismatch of derivative values with no consistency.

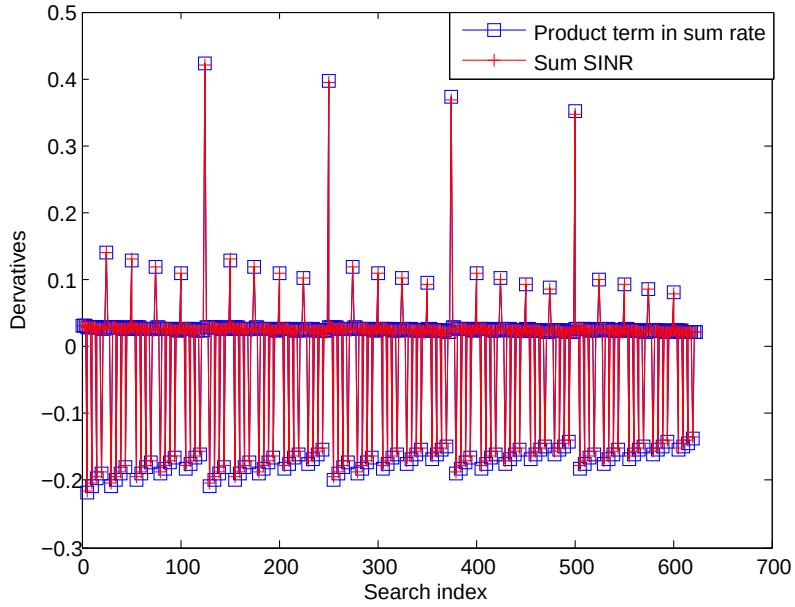

Figure B.8: One power an order of magnitude larger. Derivative values match almost perfectly.

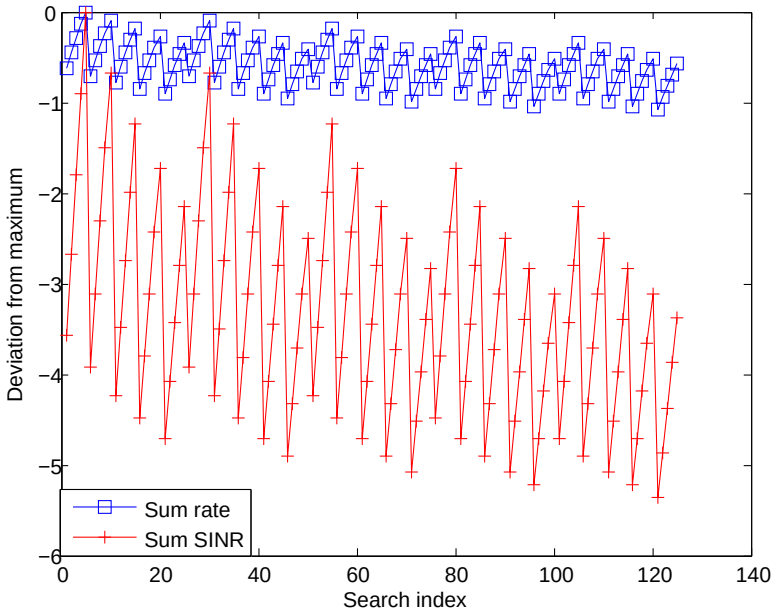

Figure B.9: Derivatives of sum rate and sum SINR with 3 transmitters including one larger power. Maxima and minima occur at the same locations, despite there being a mismatch in magnitude.

transmitters. In other words, while the logarithm does change the actual asymptotic derivative values of sum rate and sum SINR, its monotonicity ensures that the locations of maxima and minima remain the same. In the case of the chosen powers, there is one global maximum each for sum rate and sum SINR, with both occurring at the same location at search index 5.

Our simulated scenarios can exist in high load downlink HetNets, e.g., when a macro receiver receives much more power than a femto user. As shown Chapter 6, searching the vertices to maximize sum rate is much less computationally extensive for small number of users compared to conventional methods such as geometric programming, and produces near-optimal solutions. Thus, using power region vertices is a suitable near-optimal method for sum rate maximization.

B.5 Summary

We have provided a graphical and geometric description of the feasible power region for multiuser interference channels for arbitrary number of users subject to individ-

ual power and minimum user rate constraints. We have shown that sum SINR is quasiconvex with respect to any number of varying powers, and that sum SINR is an almost equivalent objective to maximize as sum rate when transmit powers are orders of magnitude apart, or when one power dominates the others. Through our findings, we confidently conclude that for multi-user interference scenarios where received powers can vary by an order of magnitude, searching for the vertices of the power region is a suitable near-optimal approach to maximizing sum rate.

Bibliography

1. CISCO, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update," Tech. Rep., 2016. (cited on pages xxiii and 3)

2. J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Select. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012. (cited on pages 1, 5, 9, and 109)

3. J. Gozalves, "Fifth-generation technologies trials [mobile radio]," *IEEE Veh. Technol. Mag.*, vol. 11, no. 2, pp. 5–13, June 2016. (cited on page 2)

4. J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest in the far east," Tech. Rep., IDC iView: IDC Anal. Future, Dec. 2012. (cited on page 2)

5. J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. K. Soong, and J. Zhang, "What will 5G be?," *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, pp. 1046–1082, June 2014. (cited on pages 2, 3, 10, 37, 46, 53, and 66)

6. F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014. (cited on page 3)

7. T. E. Bogale and L. B. Le, "Massive MIMO and mmwave for 5G wireless HetNet: Potential benefits and challenges," *IEEE Veh. Technol. Mag.*, vol. 11, no. 1, pp. 64–75, Mar. 2016. (cited on pages 3 and 109)

8. A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 10–21, June 2011. (cited on pages 3, 9, and 37)

9. H. H. Yang, J. Lee, and T. Q. S. Quek, "Heterogeneous cellular network with energy harvesting-based D2D communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1406–1419, Feb. 2016. (cited on page 5)

10. L. Wei, R.Q. Hu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: Challenges and research aspects," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 90–96, June 2014. (cited on pages 5 and 11)

11. T. Zahir, K. Arshad, A. Nakata, and K. Moessner, "Interference management in femtocells," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 293–311, First 2013. (cited on pages 6 and 19)

12. "Interference management in UMTS femtocells," Tech. Rep., Small Cell Forum Release Three, 2013. (cited on pages 6 and 19)

13. G. Boudreau, J. Panicker, Ning Guo, Rui Chang, Neng Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 74–81, Apr. 2009. (cited on page 6)

14. E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective," *IEEE Wireless Commun. Mag.*, vol. 21, no. 3, pp. 118–127, June 2014. (cited on pages 7, 9, and 53)

15. V. Chandrasekhar, J. G. Andrews, T. Muharemovic, Z. Shen, and A. Gatherer, "Power control in two-tier femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4316–4328, Aug. 2009. (cited on page 7)

16. H. B. Jung and D. K. Kim, "Power control of femtocells based on max-min fairness in heterogeneous networks," *IEEE Commun. Lett.*, vol. 17, no. 7, pp. 1372–1375, July 2013. (cited on page 7)

17. E. Castañeda, A. Silva, R. Samano-Robles, and A. Gameiro, "Distributed linear precoding and user selection in coordinated multicell systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 4887–4899, July 2016. (cited on page 7)

18. K. Wang, H. Li, F. R. Yu, W. Wei, and L. Suo, "Interference alignment in virtualized heterogeneous cellular networks with imperfect channel state information (CSI)," *IEEE Trans. Veh. Technol.*, accepted for publication, 2016. (cited on page 7)

19. D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and Wei Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Select. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010. (cited on pages 7, 19, and 108)

20. A. Wiesel, Y.C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Processing*, vol. 56, no. 9, pp. 4409–4418, Sept. 2008. (cited on pages 7, 19, 23, 24, and 30)

21. D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2016. (cited on pages 7, 8, 14, 37, 53, 65, and 109)

22. T. Zhou, Y. Huang, W. Huang, S. Li, Y. Sun, and L. Yang, "QoS-aware user association for load balancing in heterogeneous cellular networks," in *Proc. VTC-Fall*, Sept. 2014, pp. 1–5. (cited on page 8)

23. D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013. (cited on pages 8 and 65)

24. Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J.G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013. (cited on pages 8, 9, 38, 39, and 65)

25. D. Amzallag, R. Bar-Yehuda, D. Raz, and G. Scalosub, "Cell selection in 4G cellular networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1443–1455, July 2013. (cited on page 8)

26. S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *Proc. IEEE ICC*, June 2012, pp. 2457–2461. (cited on pages 8, 9, 37, 54, 60, 61, and 75)

27. W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proc. INFOCOM*, Apr. 2014, pp. 1096–1104. (cited on pages 8 and 65)

28. D. Liu, Y. Chen, K. K. Chai, and T. Zhang, "Joint uplink and downlink user association for energy-efficient hetnets using nash bargaining solution," in *Proc. IEEE VTC-Spring, 2014*, May 2014, pp. 1–5. (cited on pages 8 and 66)

29. E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 998–1006. (cited on pages 8 and 66)

30. S. Bayat, R. H. Y. Louie, Z. Han, B. Vucetic, and Y. Li, "Distributed user association and femtocell allocation in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 3027–3043, Aug. 2014. (cited on page 8)

31. T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An axiomatic theory of fairness in network resource allocation," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9. (cited on pages 11, 53, 59, and 69)

32. A. T. Gamage, H. Liang, R. Zhang, and X. Shen, "Device-to-device communication underlaying converged heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 6, pp. 98–107, Dec. 2014. (cited on pages 11 and 88)

33. D. Feng, L. Lu, Y. Yi, G. Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 49–55, Apr. 2014. (cited on page 11)

34. S. M. Mumtaz and J. Rodriguez, *Smart Device to Smart Device Communication*, Springer, 2014. (cited on pages 11 and 81)

35. J. Guo, S. Durrani, X. Zhou, and H. Yanikomeroglu, "Device-to-device communication underlaying a finite cellular network region," *IEEE Trans. Wireless Commun.*, 2016. (cited on page 11)

36. A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, Fourth quarter 2014. (cited on page 11)

37. H. Tang, Z. Ding, and B. C. Levy, "Enabling D2D communications through neighbor discovery in LTE cellular networks," *IEEE Trans. Signal Processing*, vol. 62, no. 19, pp. 5157–5170, Oct. 2014. (cited on pages 11 and 79)

38. Y. Zhao, B. Pelletier, P. Marinier, and D. Pani, "D2D neighbor discovery interference management for LTE systems," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2013, pp. 550–554. (cited on page 11)

39. C. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011. (cited on pages 12, 17, 80, 81, 82, 86, and 89)

40. G. Yu, L. Xu, D. Feng, R. Yin, G. Y. Li, and Y. Jiang, "Joint mode selection and resource allocation for device-to-device communications," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3814–3824, Nov. 2014. (cited on pages 12, 80, 81, 82, and 100)

41. Y. Huang, S. Durrani, and X. Zhou, "Interference suppression using generalized inverse precoder for downlink heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 325–328, June 2015. (cited on page 13)

42. Y. Huang, S. Durrani, and X. Zhou, "Interference nulling for offloaded heterogeneous users using macro generalized inverse precoder," in *Proc. IEEE ISCIT*, Oct. 2015. (cited on page 13)

43. Y. Huang, L. Bell, S. Durrani, X. Zhou, and N. Yang, "Effects of load dependent dynamic biasing and association order for cell range expansion," in *Proc. IEEE ICSPCS*, Dec. 2016. (cited on page 14)

44. S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013. (cited on page 15)

45. M. Mirahsan, R. Schoenen, and H. Yanikomeroglu, "HetHetNets: Heterogeneous traffic distribution in heterogeneous wireless cellular networks," *IEEE J. Select. Areas Commun.*, vol. 33, no. 10, pp. 2252–2265, Oct. 2015. (cited on pages 15 and 53)

46. Y. Huang, S. Durrani, P. Dmochowski, and X. Zhou, "A proposed network balance index for heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 6, no. 1, pp. 98–101, 2017. (cited on page 15)

47. Y. Huang, A. A. Nasir, S. Durrani, and X. Zhou, "Mode selection, resource allocation, and power control for D2D-enabled two-tier cellular network," *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3534–3547, Aug. 2016. (cited on page 17)

48. Y. Huang, A. A. Nasir, S. Durrani, and X. Zhou, "Graphical generalization of power control in multiuser interference channels," in *Proc. IEEE AusCTW*, Jan. 2016. (cited on page 17)

49. S. Sun, Q. Gao, Yg Peng, Yingmin Wang, and Lingyang Song, "Interference management through CoMP in 3GPP LTE-advanced networks," *IEEE Wireless Commun. Mag.*, vol. 20, no. 1, pp. 59–66, Feb. 2013. (cited on page 19)

50. M. Cierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting, "On number of almost blank subframes in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 5061–5073, Oct. 2013. (cited on page 19)

51. W. Nam, D. Bai, J. Lee, and I. Kang, "Advanced interference management for 5G cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 52–60, May 2014. (cited on page 19)

52. D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Q. S. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heteroge-

neous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 22–30, June 2011. (cited on page 19)

53. S. Cheng, S. Lien, F. Chu, and K. Chen, "On exploiting cognitive radio to mitigate interference in macro/femto heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 40–47, June 2011. (cited on page 19)

54. Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 60–67, Oct. 2004. (cited on page 20)

55. V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the *k*-user interference channel," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008. (cited on page 20)

56. A.R. Elsherif, Z Ding, and X. Liu, "Dynamic MIMO precoding for femtocell interference mitigation," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 648–666, Feb. 2014. (cited on page 20)

57. Z. Xu, C. Yang, G.Ye. Li, Y. Liu, and S. Xu, "Energy-efficient CoMP precoding in heterogeneous networks," *IEEE Trans. Signal Processing*, vol. 62, no. 4, pp. 1005–1017, Feb. 2014. (cited on page 20)

58. A. Manolakos, Y. Noam, and A. J. Goldsmith, "Null space learning in cooperative MIMO cellular networks using interference feedback," in *Proc. IEEE GLOBECOM*, Dec. 2013. (cited on page 23)

59. 3GPP, "Simulation assumptions and parameters for FDD HeNB RF requirements," Tech. Rep., May 2009. (cited on pages 23, 32, and 34)

60. P. C. Hansen, "Regularization tools - a Matlab package for analysis and solution of discrete ill-posed problems - version 3.0 for Matlab 5.2," *Numer. Algorithms*, vol. 46, pp. 189–194, 2007. (cited on pages 28, 30, and 115)

61. E. Jury and M. Mansour, "Positivity and nonnegativity conditions of a quartic equation and related problems," *IEEE Trans. Automat. Contr.*, vol. 26, no. 2, pp. 444–451, Apr. 1981. (cited on page 31)

62. K. Kikuchi and H. Otsuka, "Proposal of adaptive control CRE in heterogeneous networks," in *Proc. IEEE PIMRC*, Sept. 2012, pp. 910–914. (cited on page 37)

63. Y. Wang, S. Chen, H. Ji, and H. Zhang, "Load-aware dynamic biasing cell association in small cell networks," in *Proc. IEEE ICC*, June 2014, pp. 2684–2689. (cited on pages 37, 39, and 44)

64. S. Sun, W. Liao, and W. Chen, "Traffic offloading with rate-based cell range expansion offsets in heterogeneous networks," in *Proc. IEEE WCNC*, Apr. 2014, pp. 2833–2838. (cited on page 38)

65. R. Han, C. Feng, and H. Xia, "Optimal user association based on topological potential in heterogeneous networks," in *Proc. IEEE PIMRC*, Sept. 2013, pp. 2409–2413. (cited on page 38)

66. T. Zhou, Y. Huang, W. Huang, S. Li, Y. Sun, and L. Yang, "QoS-aware user association for load balancing in heterogeneous cellular networks," in *Proc. IEEE VTC-Fall*, Sept. 2014, pp. 1–5. (cited on pages 38 and 65)

67. C. Zhao and C. Hua, "Traffic-load aware user association in dense unsaturated wireless networks," in *Proc. WCSP*, Oct. 2014, pp. 1–6. (cited on page 38)

68. H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012. (cited on page 45)

69. H. Kim, G. Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Networking*, vol. 20, no. 1, pp. 177–190, Feb. 2012. (cited on page 53)

70. A. Okabe, B. Boots, and K. Sugihara, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, John Wiley & Sons, Inc., New York, NY, USA, 1992. (cited on page 56)

71. R. Pure and S. Durrani, "Computing exact closed-form distance distributions in arbitrarily-shaped polygons with arbitrary reference point," *The Mathematica Journal*, vol. 17, June 2015. (cited on page 56)

72. R. K. Ganti and M. Haenggi, "Regularity, interference, and capacity of large ad hoc networks," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Oct. 2006, pp. 3–7. (cited on page 60)

73. D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *American Mathematical Monthly*, pp. 9–15, Jan. 1962. (cited on pages 65 and 68)

74. X. Tang, P. Ren, Y. Wang, Q. Du, and L. Sun, "User association as a stochastic game for enhanced performance in heterogeneous networks," in *Proc. IEEE ICC*, June 2015, pp. 3417–3422. (cited on page 65)

75. B. Rengarajan and G. de Veciana, "Practical adaptive user association policies for wireless systems with dynamic interference," *IEEE/ACM Trans. Networking*, vol. 19, no. 6, pp. 1690–1703, Dec. 2011. (cited on page 66)

76. D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1835–1850, Mar. 2016. (cited on page 66)

77. J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 803–826, Second quarter 2015. (cited on page 66)

78. H. Yanikomeroglu, E. Kalantari, and A. Yongacoglu, "On the number and 3D placement of drone base stations in wireless cellular networks," in *Proc. IEEE VTC-Fall*, Sept. 2016. (cited on page 66)

79. A.F. Beardon, "Sum of powers of integers," *American Mathematical Monthly*, pp. 201–213, Mar. 1996. (cited on page 70)

80. F. Malandrino, C. Casetti, and C.-F. Chiasserini, "Toward D2D-enhanced heterogeneous networks," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 94–100, Nov. 2014. (cited on page 79)

81. X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014. (cited on pages 79 and 81)

82. H. ElSawy, E. Hossain, and M.-S. Alouini, "Analytical modeling of mode selection and power control for underlay D2D communication in cellular networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4147–4161, Nov. 2014. (cited on pages 79 and 81)

83. K. Doppler, C.Yu, C. B. Ribeiro, and P. Janis, "Mode selection for device-to-device communication underlaying an LTE-Advanced network," in *Proc. IEEE WCNC*, Apr. 2010. (cited on pages 79, 81, and 82)

84. H. Min, J. Lee, S. Park, and D. Hong, "Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 3995–4000, Dec. 2011. (cited on pages 79, 81, and 82)

85. D. Feng, G. Yu, C. Xiong, Y. Yi, G. Ye Li, G. Feng, and S. Li, "Mode switching for energy-efficient device-to-device communications in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6993–7003, Dec. 2015. (cited on pages 80 and 96)

86. H. Tang and Z. Ding, "Mixed mode transmission and resource allocation for D2D communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 162–175, Jan. 2016. (cited on page 80)

87. D. Feng, L. Lu, Y. Yi, G. Y. Li, G. Feng, and S. Li, "Device-to-device communications underlaying cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3541–3551, Aug. 2013. (cited on pages 80, 81, and 89)

88. A. Gjendemsj, D. Gesbert, G. E. Oien, and S. G. Kiani, "Binary power control for sum rate maximization over multiple interfering links," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3164–3173, Aug. 2008. (cited on pages 80, 81, 89, 121, 122, and 125)

89. W. Zhong, Y. Fang, S. Jin, K. Wong, S. Zhong, and Z. Qian, "Joint resource allocation for device-to-device communications underlaying uplink MIMO cellular networks," *IEEE J. Select. Areas Commun.*, vol. 33, no. 1, pp. 41–54, Jan. 2015. (cited on pages 80 and 81)

90. H. Zhang, C. Jiang, X. Mao, and H. Chen, "Interference-limited resource optimization in cognitive femtocells with fairness and imperfect spectrum sensing," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1761–1771, Mar. 2016. (cited on page 80)

91. F. Wang, C. Xu, L. Song, and Z. Han, "Energy-efficient resource allocation for device-to-device underlay communication," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2082–2092, Apr. 2015. (cited on pages 80 and 81)

92. Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 441–454, Feb. 2015. (cited on pages 80 and 81)

93. R. Yin, C. Zhong, G. Yu, Z. Zhang, K. K. Wong, and X. Chen, "Joint spectrum and power allocation for D2D communications underlaying cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2182–2195, Apr. 2016. (cited on page 80)

94. R. Yin, G. Yu, H. Zhang, Z. Zhang, and G. Y. Li, "Pricing-based interference coordination for D2D communications in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1519–1532, Mar. 2015. (cited on page 80)

95. H. Zhang, C. Jiang, N.C. Beaulieu, X. Chu, X. Wang, and T. Q. S. Quek, "Resource allocation for cognitive small cell networks: A cooperative bargaining

game theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3481–3493, June 2015. (cited on page 80)

96. H. Song, J. Y. Ryu, W. Choi, and R. Schober, "Joint power and rate control for device-to-device communications in cellular systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5750–5762, Oct. 2015. (cited on page 81)

97. D. Zhu, J. Wang, A. L. Swindlehurst, and C. Zhao, "Downlink resource reuse for device-to-device communications underlaying cellular networks," *IEEE Signal Processing Lett.*, vol. 21, no. 5, pp. 531–534, May 2014. (cited on page 81)

98. X. Chen, L. Chen, M. Zeng, X. Zhang, and D. Yang, "Downlink resource allocation for device-to-device communication underlaying cellular networks," in *Proc. IEEE PIMRC*, Sept. 2012, pp. 232–237. (cited on page 81)

99. F. Malandrino, C. Casetti, C. F. Chiasserini, and Z. Limani, "Uplink and downlink resource allocation in D2D-enabled heterogeneous networks," in *Proc. IEEE WCNC Workshop*, Apr. 2014, pp. 87–92. (cited on page 81)

100. X. Lin, J. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014. (cited on page 81)

101. T. D. Novlan and J. G. Andrews, "Analytical evaluation of uplink fractional frequency reuse," *IEEE Trans. Commun.*, vol. 61, no. 5, pp. 2098–2108, May 2013. (cited on page 81)

102. X. Ma, J. Liu, and H. Jiang, "Resource allocation for heterogeneous applications with device-to-device communication underlaying cellular networks," *IEEE J. Select. Areas Commun.*, vol. 34, no. 1, pp. 15–26, Jan. 2016. (cited on page 82)

103. M. Chiang, C.W. Tan, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, July 2007. (cited on pages 90, 103, and 121)

104. W. Hardjawana, B. Vucetic, Y. Li, and Z. Zhou, "Spectrally efficient wireless systems with cooperative precoding and beamforming," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5871–5882, Dec. 2009. (cited on page 108)

105. L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014. (cited on page 109)

106. C. Mollén, E. G. Larsson, and T. Eriksson, "Waveforms for the massive MIMO downlink: Amplifier efficiency, distortion, and performance," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5050–5063, Dec. 2016. (cited on page 109)

107. F. Boccardi, J. Andrews, H. Elshaer, M. Dohler, S. Parkvall, P. Popovski, and S. Singh, "Why to decouple the uplink and downlink in cellular networks and how to do it," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 110–117, Mar. 2016. (cited on page 109)

108. H. Song, X. Fang, and L. Yan, "Handover scheme for 5G C/U plane split heterogeneous network in high-speed railway," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4633–4646, Nov. 2014. (cited on page 109)

109. J. K. Baksalary and O. M. Baksalary, "Particular formulae for the Moore–Penrose inverse of a columnwise partitioned matrix," *Linear Algebra and its Applications*, vol. 421, no. 1, pp. 16 – 23, 2007. (cited on page 111)

110. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004. (cited on page 116)

111. K. Illanko, A. Anpalagan, E. Hossain, and D. Androutsos, "On the power allocation problem in the gaussian interference channel with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 1101–1115, Feb. 2014. (cited on page 121)

112. K. Illanko, A. Anpalagan, and D. Androutsos, "Convex structure of the sum rate on the boundary of the feasible set for coexisting radios," in *Proc. IEEE ICC*, June 2011, pp. 1–6. (cited on page 121)

113. T. Holliday, A. Goldsmith, P. Glynn, and N. Bambos, "Distributed power and admission control for time varying wireless networks," in *Proc. IEEE GLOBECOM*, Nov. 2004, vol. 2, pp. 768–774 Vol.2. (cited on page 121)

114. L. Qian, Y. Zhang, and J. Huang, "MAPEL: Achieving global optimality for a non-convex wireless power control problem," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1553–1563, Mar. 2009. (cited on page 121)

115. S.A. Jafar and A. Goldsmith, "Adaptive multirate CDMA for uplink throughput maximization," *IEEE Trans. Wireless Commun.*, vol. 2, no. 2, pp. 218–228, Mar. 2003. (cited on page 122)

116. H. Boche, S. Naik, and T. Alpcan, "Characterization of convex and concave resource allocation problems in interference coupled wireless systems," *IEEE Trans. Signal Processing*, vol. 59, no. 5, pp. 2382–2394, May 2011. (cited on page 122)