The Dissertation Committee for Qiaoyang Ye
certifies that this is the approved version of the following dissertation:

Small Cell and D2D Offloading in Heterogeneous Cellular Networks

Committee:

Jeffrey G. Andrews, Co-supervisor

Constantine Caramanis, Co-supervisor

Francois Baccelli

David Morton

Sanjay Shakkottai

Sriram Vishwanath

Small Cell and D2D Offloading in Heterogeneous Cellular Networks

by

Qiaoyang Ye, B.E.; M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

Dedicated to my parents.

Acknowledgments

My years as a graduate student has been a wonderful journey, in which I am greatly indebted to many people for their help and support.

Foremost, I would like to express my deepest gratitude to my supervisors Prof. Jeffrey G. Andrews and Prof. Constantine Caramanis, for their continuous support and guidance of my study and research at UT. I benefited a lot from Jeff's keenness in research directions, great skills in writing and presentation, and his great personality and enthusiasm, which profoundly inspired me. I also have been truly fortunate to have Constantine as my supervisor. His strong technical strength, passion in research, and guidance with great patience have immensely motivated me. Their trust and dedication were fundamental to all my accomplishments at UT. I could not have imagined having better advisors and mentors for my Ph.D. study. Also, I would like to thank Prof. Francois Baccelli, Prof. David Morton, Prof. Sanjay Shakkottai and Prof. Sriram Vishwanath for serving on my thesis committee and providing me with invaluable suggestions and helpful comments on my research works.

I gratefully acknowledge Huawei and National Science Foundation for supporting my research. I would like to express my special gratitude to Mr. Mazin Al-Shalash for hosting me at Huawei, Plano, TX, during summer 2012 and summer 2013. These internships as well as our collaborative research

projects have motivated my research on load balancing in heterogeneous networks and the resource allocation of device-to-device communication. My sincere thanks also goes to Dr. Ozgun Y. Bursalioglu and Dr. Haralabos C. Papadopoulos for hosting me at Docomo Innovations Inc., Palo Alto, CA, during summer 2014. This internship inspired my work on load balancing in massive MIMO heterogeneous networks. The knowledge and wisdom I learned from these industry collaborators are tremendously helpful in my research, and will remain beneficial for my future career.

My life at UT would never have been so memorable without the brilliant colleagues and friends. I thank Arthur, Beiyu, Chun-Hung, Derya, Harpreet, Mandar, Ping, Sarabjot, Xingqin and Yudong for various technical discussions. A special thanks to Xingqin for sharing with me his enormous experiences and being a inspiring colleague. I would also like to thank my Chinese friends for sharing the most enjoyable time with me: Beiyu, Chao, Hongbo, Jing, Jingwen, Ping, Song, Tianyang, Yingxi, Yingzhe, Yudong, Yuhao, Yuhuan, Xingqin, Zheng and many others. I would also like to thank Melanie Gulick from ECE graduate office and Janet Preuss, Karen Little, and Lauren Bringle from WNCG office for their help in administrative work.

Last, but definitely not least, I would like to thank my parents, Zhiyuan and Zhuanghong, for their most generous love and endless support throughout my life. I also thank Jialin for always being there cheering me up.

Small Cell and D2D Offloading in Heterogeneous Cellular Networks

Qiaoyang Ye, Ph.D.
The University of Texas at Austin, 2015

Co-supervisors: Jeffrey G. Andrews
Constantine Caramanis

Future wireless networks are evolving to become ever more heterogeneous, including small cells such as picocells and femtocells, and direct device-to-device (D2D) communication that bypasses base stations (BSs) altogether to share stored and personalized content. Conventional user association schemes are unsuitable for heterogeneous networks (HetNets), due to the massive disparities in transmit power and capabilities of different BSs. To make the most of the new low-power infrastructure and D2D communication, it is desirable to facilitate and encourage users to be offloaded from the macro BSs. This dissertation characterizes the gain in network performance (e.g., the rate distribution) from offloading users to small cells and the D2D network, and develops efficient user association, resource allocation, and interference management schemes aiming to achieve the performance gain.

First, we optimize the load-aware user association in HetNets with single-antenna BSs, which bridges the gap between the optimal solution and

a simple small cell biasing approach. We then develop a low-complexity distributed algorithm that converges to a near-optimal solution with a theoretical performance guarantee. Simulation results show that the biasing approach loses surprisingly little with appropriate bias factors, and there is a large rate gain for cell-edge users.

This framework is then extended to a joint optimization of user association and resource blanking at the macro BSs – similar to the enhanced intercell interference coordination (eICIC) proposed in the global cellular standards, 3rd Generation Partnership Project (3GPP). Though the joint problem is nominally combinatorial, by allowing users to associate to multiple BSs, the problem becomes convex. We show both theoretically and through simulation that the optimal solution of the relaxed problem still results in a mostly unique association. Simulation shows that resource blanking can further improve the network performance.

Next, the above framework with single-antenna transmission is extended to HetNets with BSs equipped with large-antenna arrays and operating in the massive MIMO regime. MIMO techniques enable the option of another interference management: serving users simultaneously by multiple BSs – termed *joint transmission* (JT). This chapter formulates a unified utility maximization problem to optimize user association with JT and resource blanking, exploring which an efficient dual subgradient based algorithm approaching optimal solutions is developed. Moreover, a simple scheduling scheme is developed to implement near-optimal solutions.

We then change direction slightly to develop a flexible and tractable framework for D2D communication in the context of a cellular network. The model is applied to study both shared and orthogonal resource allocation between D2D and cellular networks. Analytical SINR distributions and average rates are derived and applied to maximize the total throughput, under an assumption of interference randomization via time and/or frequency hopping, which can be viewed as an optimized lower bound to other more sophisticated scheduling schemes.

Finally, motivated by the benefits of cochannel D2D links, this dissertation investigates interference management for D2D links sharing cellular uplink resources. Showing that the problem of maximizing network throughput while guaranteeing the service of cellular users is non-convex and hence intractable, a distributed approach that is computationally efficient with minimal coordination is proposed instead. The key algorithmic idea is a pricing mechanism, whereby BSs optimize and transmit a signal depending on the interference to D2D links, who then play a best response (i.e., selfishly) to this signal. Numerical results show that our algorithms converge quickly, have low overhead, and achieve a significant throughput gain, while maintaining the quality of cellular links at a predefined service level.

Table of Contents

Acknowledgments **v**

Abstract **vii**

List of Tables **xv**

List of Figures **xvi**

Chapter 1. Introduction **1**
1.1 Ongoing Evolution to HetNets 2
1.2 Load Balancing and Interference Management 4
1.2.1 The Need for Load-aware Association 4
1.2.2 Interference Management 6
1.3 Contributions and Organization 9

Chapter 2. Load Balancing in HetNets with Small BSs **15**
2.1 Related Work 16
2.2 Contributions and Organization 18
2.3 System Model 19
2.4 Problem Formulation 22
2.4.1 General Utility Maximization: Unique Association . . . 22
2.4.2 General Utility Maximization: Allowing Joint Association 23
2.4.3 Logarithmic Utility Formulation 24
2.4.4 Analysis of Optimized Resource Allocation 25
2.4.5 Relaxation to Fractional User Association 28
2.5 The Distributed Algorithm Based on the Dual Subgradient Method 29
2.5.1 Dual Decomposition 30
2.5.2 The Distributed Algorithm 31
2.5.3 Step Size and Convergence 34

2.6 Cell Range Expansion (Biasing) 35
2.6.1 SINR Bias 37
2.6.2 Rate Bias 37
2.7 Performance Evaluation 38
2.7.1 Loads among different BSs 39
2.7.2 Rate CDF 40
2.7.3 Biasing Factor 42
2.8 Summary 45

Chapter 3. Joint Optimization of User Association and Resource Blanking 48
3.1 Related Work 49
3.2 Contributions 50
3.3 System Model 51
3.4 Problem Formulation 52
3.5 Performance Evaluation 59
3.6 Summary 66

Chapter 4. User Association and Interference Management in Massive MIMO HetNets 68
4.1 Related Work 69
4.2 Contributions and Organization 74
4.3 System Model 76
4.3.1 Channel Model 76
4.3.2 Admissible Joint Transmission 77
4.3.3 Admissible Transmission with blanking 81
4.4 Instantaneous Rate and Long-term Rate 82
4.4.1 Instantaneous rate 83
4.4.2 Long-term Rate 84
4.5 Unified NUM Problem Formulation 86
4.5.1 Restricting Options of ATMs 86
4.5.2 The Unified NUM Problem 87
4.6 Dual Subgradient Based Algorithm 90

4.6.1 The Dual Subgradient Method . . . 91
4.6.2 Finding the Optimal Primal Solutions Given Optimal Lagrangian Multipliers . . . 92
4.6.3 Implementation Discussion . . . 94
4.7 Virtual Queue Based Scheduling Scheme . . . 97
4.7.1 The Feasibility of the NUM Solution in Implementation 97
4.7.2 The Greedy Virtual Queue Scheduling Scheme . . . 98
4.8 Possible Alterantive Method – the Slot-based Gradient Algorithm [1] . . . 101
4.9 Performance Evaluation . . . 103
4.10 Summary . . . 114
4.11 Appendix . . . 115
4.11.1 Proof of Spectral Efficiency Using ZF Precoding . . . 115
4.11.2 Proof of Spectral Efficiency Using MRT Precoding . . . 116
4.11.3 Proof of Theorem 4.1 . . . 117
4.11.4 Proof of Proposition 4.5 . . . 119

Chapter 5. Analysis and Optimization of D2D Enhanced Cellular Networks Using Time-Frequency Hopping 121
5.1 Related Work . . . 122
5.2 Contributions . . . 124
5.3 System Model . . . 126
5.3.1 Deployment of D2D and Cellular Networks . . . 126
5.3.2 Scheduling Scheme . . . 128
5.3.3 Load Modeling . . . 130
5.3.4 Channel Model . . . 132
5.4 Analysis of the Dedicated Network . . . 132
5.4.1 SINR Distribution . . . 132
5.4.2 Rate Analysis in the Dedicated Network . . . 137
5.5 Analysis of the Shared Network . . . 139
5.5.1 SINR Distribution of D2D links . . . 139
5.5.2 SINR Distribution of Cellular Users . . . 141
5.5.3 Rate Analysis in the Shared Network . . . 142

5.6 Optimization of the D2D-Cellular Network 144
5.6.1 Optimization of the Dedicated Network 145
5.6.2 Optimization of the Shared Network 149
5.7 Performance Evaluation 150
5.7.1 Validation of the System Model 150
5.7.2 Optimization of Network Performance 154
5.8 Summary 160
5.9 Appendix 161
5.9.1 Proof of Proposition 5.1 161
5.9.2 Proof of Corollary 5.2 162
5.9.3 Proof of Proposition 5.5 162
5.9.4 Proof of Proposition 5.6 163
5.9.5 Proof of Proposition 5.7 164

Chapter 6. Distributed Resource Allocation in D2D Enhanced Cellular Networks 166

6.1 Related Work 167
6.2 Contributions 169
6.3 System Model 171
6.4 Problem Formulation 174
6.4.1 Single-stage Problem Formulation 174
6.4.2 Two-stage Problem Formulation 176
6.5 Lower Problem: A Non-cooperative D2D Network 178
6.5.1 Distributed Algorithm Design 179
6.5.2 Joint Resource Allocation and Power Control – A Lower Bound Problem 181
6.5.3 Algorithm Design for the Lower bound Problem 184
6.6 Upper Problem: Network's Pricing Mechanism 189
6.6.1 An Equivalent Upper Problem 190
6.6.2 Algorithm Design for the Upper Problem 191
6.7 Performance Evaluation 196
6.7.1 The Lower Problem: D2D Non-cooperative Game 197
6.7.2 The Upper Problem: Network Pricing Mechanism . . . 198

6.8 Summary 205
6.9 Appendix 207
6.9.1 Proof of Proposition 6.3 207
6.9.2 Proof of Proposition 6.5 208
6.9.3 Proof of Proposition 6.6 210

Chapter 7. Conclusions **211**
7.1 Summary 211
7.2 Future Directions 214

Bibliography **217**

Vita **246**

List of Tables

3.1 Simulation parameters for user association and RB blanking optimization 61

4.1 Example of RBs enabled by admissible transmission schemes over 4 BSs. 79

5.1 Notation summary for D2D-cellular networks with time-frequency hopping 133

5.2 Simulation parameters for D2D-cellular networks with time-frequency hopping 151

List of Figures

1.1 Illustration of a HetNet including macro BSs, picocells, femtocells and D2D communication. 4

1.2 Max-SINR association versus load-aware association. Lines indicate the user association and highlight the differences in cell association policies. 7

2.1 Comparisons of average number of users per tier in a three-tier HetNet. 40

2.2 The CDFs of overall rate in a three-tier HetNets, in both static setting and stochastic setting. The biasing factors of macro BSs, picos and femtos are $\{A_1, A_2, A_3\} = \{1.00, 4.00, 11.9\}$ in SINR bias, and $\{B_1, B_2, B_3\} = \{1.00, 1.59, 1.88\}$ in rate bias, respectively. 41

2.3 The CDFs of overall rate in macro-only networks. 42

2.4 Rate gain in a three-tier HetNet. The rate ratio of joint association scheme, fractional-rounding scheme, dual distributed algorithm, and CRE to max-SINR association is represented. 43

2.5 Biasing factors vs. density of small BSs in a three-tier HetNet. The density of one tier changes while the others are fixed, with tier 1 always having biasing factor 1. 44

2.6 Biasing factors vs. transmit power of BSs in a three-tier HetNet, with tier 1 having biasing factor 1. 46

3.1 Examples of graph representation which illustrates the fractional user association. 56

3.2 Examples of associations in HetNets with different association schemes. The dashed lines indicate the association in blank RBs, while the solid lines indicate the association in normal RBs. 60

3.3 The rate distribution of users with different association schemes. 62

3.4 The average of optimal fraction of blank RBs (i.e., z) vs. density of small cells. 63

3.5 Load versus small cell density in a two-tier HetNet. 64

3.6 The throughput gain of worst 10% users in networks with different deployments of small cells. 66

4.1 Illustration of instantaneous rate versus user locations, when $S_j(|\mathcal{C}|) = |\mathcal{C}|S_j$. The location in Fig. 4.1b indicates the x-axis coordinate of the path in Fig. 4.1a. 85

4.2 The illustration of network deployment. The white grids are the regular areas, while the shadowed grids are hotspots. . . . 105

4.3 The geometric mean of rates using different approaches, when $\rho = 1$. 106

4.4 The geometric mean of rates with RB blanking, when $\rho = 1$ and macro and pico BSs share the resources. Blanking further improves the network performance. 107

4.5 The comparison between the geometric mean of long-term rates using the slot-based framework with greedy algorithm and the greedy VQ scheme, when $\rho = 1$ in cases where macro and pico BSs share resources. The utility obtained from the greedy algorithm in the slot-based framework is much less the utility obtained from the greedy VQ scheme. 108

4.6 The long-term rate CDF using different approaches, when $\rho = 1$. 109

4.7 The rate differences between the greedy VQ scheduling scheme and the NUM solution, when $\rho = 1$ and macro and pico BSs share resources. There are 82.62% users whose rate differences between the greedy VQ scheduling scheme and the NUM solution are within 10% of the NUM solution. 110

4.8 The number of users served by different clusters, when $\rho = 1$. The "Cluster UEs" refer to the users served by clusters of size larger than 1. Most users have unique association. 111

4.9 The fraction of resources allocated by BSs to different clusters, when $\rho = 1$. There are 2 and 3 active ATMs in Figs. 4.9a and 4.9b, respectively. 112

4.10 The geometric mean of rates using different approaches versus ρ. As ρ decreases, the gain from JT decreases in both shared and orthogonal operation scenarios. 112

4.11 The fraction of resources allocated to clusters of different sizes with blanking. As ρ decreases, more resources are allocated to clusters of smaller size. 113

5.1 Illustration of the network model. The red points are BSs which are deployed according to a PPP. The D2D links include both silent potential D2D links (with dashed lines) and active D2D links (with solid lines). 128

5.2 Illustration of the time-frequency hopping scheme. The shadowed squares are the RBs occupied by some active D2D links. A potential D2D link accesses each time slot uniformly with probability p_t and accesses each subband uniformly with probability p_f. 130

5.3 The SINR CDFs of active D2D links and cellular users in the dedicated network, with hopping probabilities $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.2$ and $p_{f_2} = 0.6$. 152

5.4 The SINR CDFs of D2D links and cellular users in the shared network. The hopping probabilities are $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.1$ and $p_{f_2} = 0.3$. 153

5.5 The average rates vs. the density of users in the dedicated network ($\theta = 0.5$). The hopping probabilities are $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.2$ and $p_{f_2} = 0.6$. The density of potential D2D links increases proportionally to the density of cellular users. The dashed lines are the simulation results while the solid lines are the corresponding analytical results. 154

5.6 The average rates vs. the density of users in the shared network. The hopping probabilities are $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.1$ and $p_{f_2} = 0.3$. The dashed lines are the simulation results while the solid lines are the corresponding analytical results. 155

5.7 Effect of time hopping probabilities on the total rate density in heavily loaded networks ($\theta = 0.5$). The solid curves and dashed curves show the performance of exact rates and their lower bounds, respectively. We let frequency hopping probabilities be $p_{f_i} = \min\{1, B_C/b_{D_i}\}$. 156

5.8 Effect of frequency hopping probabilities on the total rate density in heavily loaded networks ($\theta = 0.5$). The solid curves and dashed curves show the performance of exact rates and their lower bounds, respectively. We let the time hopping probabilities be $p_{t_i} = 1$. 157

5.9 Total rate versus θ. The solid curves and dashed curves show the performance of exact rates and their lower bounds, respectively. We let frequency hopping probabilities be $p_{f_i} = \min\{1, B_C/b_{D_i}\}$ and the time hopping probabilities be $p_{t_i} = 1$. 158

5.10 Rate versus θ in a network with the average distance between a D2D transmitter and its receiver being 280m. We let frequency hopping probabilities be $p_{f_i} = \min\{1, B_C/b_{D_i}\}$ and the time hopping probabilities be $p_{t_i} = 1$. 159

5.11 Effect of parameter w on the optimal mode selection to maximize the total rate. 159

6.1 Illustration of the proposed algorithm. The arrows filled with dark color indicate the procedures requiring message exchange, while the arrows filled with light color indicate the procedures involving only local measurements. The lower part describes the LB Algorithm for the lower problem, while the upper part illustrates algorithms proposed for the upper problem. 185

6.2 The access probabilities of D2D links vs. μ. The areas in the dark shade show the locations of silent D2D links with $x_i = 0$. The light shaded areas show the locations of active D2D links (i.e., $x_i \in (0, 1)$). The remaining parts show the locations of saturated D2D links (i.e., $x_i = 1$). 189

6.3 The rate distributions of D2D and cellular links using different algorithms in a single-cell network. 199

6.4 The convergence of different algorithms. 199

6.5 The total rates of cellular and D2D links using different approaches. 200

6.6 The rate distribution of cellular links using different approaches. 202

6.7 The rate distribution of D2D links using different approaches. 203

6.8 The rate of cellular links vs. the interference tolerance level. The normalized interference tolerance level means that the interference tolerance level Q is divided by the signal of the cellular link accessing the considered RB. 204

6.9 The total rate of D2D links vs. the interference tolerance level. The normalized interference tolerance level means that the interference tolerance level Q is divided by the signal of the cellular link accessing the considered RB. 204

6.10 The total rate of cellular links vs. different D2D densities. . . 205

6.11 The total rate of D2D links vs. different D2D densities. 206

Chapter 1

Introduction

Explosive mobile data traffic growth has caused cellular networks to evolve into dense and irregular heterogeneous cellular networks, especially through the proliferation of small base stations (BSs) underlaid in a conventional (macrocell) network. These small BSs include microcells, picocells and femtocells, which differ primarily in maximum transmit power, physical size, and ease-of-deployment [2]. Besides small cells, device-to-device (D2D) communication – which allows devices to directly communicate with each other – is also emerging as an important technology component for future wireless communication networks [3]. Such a paradigm shift poses many new challenges in the network design, and necessitates a significant rethinking of user association, resource allocation and interference management approaches [4] .

We start this introductory chapter with a description of heterogeneous networks (HetNets) that consist of macro BSs, small cells and D2D communication. We then discuss the necessarity and importance of load balancing and interference management in Section 1.2. Finally, Section 1.3 summarizes the main contributions of this dissertation along with its organization.

1.1 Ongoing Evolution to HetNets

Drivers of small BS deployment. The wireless traffic pattern is dramatically changing with the proliferation of ubiquitous mobile devices such as smartphones, tablets, and new "wearable" devices (e.g., smart watches), along with the invention of various data-hungry applications and services. It is expected to have a 10-fold mobile traffic growth by 2019 [5, 6], on the top of a 10-fold increase in last 5 years. Such stunning traffic growth leads to an urgent need of data rate improvement.

In current wireless systems with fixed or nearly fixed spectral resources, improvement in spectrum utilization efficiency is quite saturated, since the spectrum efficiency of a point-to-point link is very close to the theoretical limit with mature physical layer techniques [7, 8]. On the other hand, releasing more spectrum is a costly solution, and the scarcity of current ultra high frequency (UHF) bands (i.e., frequencies between 300MHz and 3GHz) results in a shortfall in boosting capacity. Though recently there is considerable interest in millimeter wave (spectrum above 30GHz) with gigahertz of spectrum available, it is likely to take more than ten years before the wide commercialization of such a novel technique [4]. Thus, one of the most promising solutions for the current capacity crunch is to deploy small cells such as picocells and femtocells to increase the cell density. In this dissertation, we let each tier model a particular type of BSs. For example, tier 1 refers to the tower-mounted macro BSs which have the largest transmit power, and tier 2 and tier 3 could be interpreted as pico and femto BSs, respectively. In gen-

eral, pico BSs transmit at a much lower power with a higher deployed density than macro BSs, while femto BSs may be deployed very densely but have the smallest transmit power.

Drivers of D2D communication. The mobile revolution through increasing mobile devices has also propelled various proximity-based services, such as content sharing, multiplayer gaming, social networking services and mobile advertising [3,9]. The existing techniques enabling such services can be broadly categorized into peer-to-peer (P2P) communication and over-the-top (OTT) solution (with a server located in the cloud) [10], with the disadvantages of uncontrolled interference in unlicensed band (e.g., WiFi Direct [11] as a P2P example) and/or low energy efficiency (e.g., Highlight App as an OTT example). Unlike the aforementioned techniques or general ad-hoc networks, D2D communication is proposed as a promising technique to meet the surging proximity-based service demand, which can operate on licensed bands and benefit from cellular infrastructure (e.g., network coordinated device discovery, synchronization and enhanced security) [3, 12–15]. Taking the advantage of physical proximity of communicating devices, D2D communication increases area spectrum efficiency, reduces energy consumption, and more importantly, it can be adopted as an offloading approach for cellular networks, leading to better resource utilization.

In summary, the exploding wireless traffic demand has led to a network evolution towards integration of small BSs and D2D communication into conventional macrocellular networks – termed HetNets in this dissertation, as

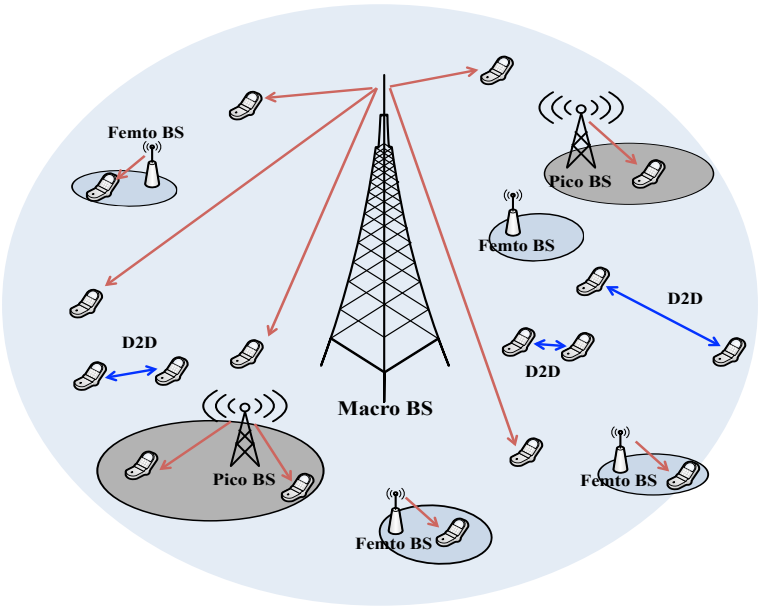


Figure 1.1: Illustration of a HetNet including macro BSs, picocells, femtocells and D2D communication.

illustrated in Fig. 1.1. HetNets introduce fundamental changes in network design and analysis. In this dissertation, we focus on the offloading and resource allocation aspects as discussed in the following section.

1.2 Load Balancing and Interference Management

1.2.1 The Need for Load-aware Association

Conventional macrocellular networks consist of homogeneous macro BSs with almost the same transmit power and regular deployment. In such macrocellular networks, a natural association scheme that decides which users should associate to which BSs is to let users connect to the BSs providing the largest signal-to-interference-plus-noise ratio (SINR) – called *max-SINR*

association, using which the number of users per macro BS is about the same if users are uniformly distributed. Hence, the need for further load balancing is usually not significant.

However, in a network with small BSs, there are massive disparities in transmit power and capability of different BSs, and thus the conventional user association schemes are no longer suitable. As an example, Fig. 1.2a shows the user association based on the max-SINR rule, which associates most users to high-power macro BSs and thus leads to a major load imbalance. Such load imbalance would exist even in HetNets with a targeted deployment, where small BSs are deployed in hotspots (i.e., high-traffic zone). Though the SINR is maximized in this case, users have to share resources with others in the same cell, and the limited resources per user in congested macro cells would result in a small overall throughput, while the resources at small BSs are significantly underutilized. Moreover, the irregular deployment of small BSs and the non-uniform user distribution further demand the rethinking of user association schemes. The critical missing piece in conventional user association schemes is the load of BSs, which provides a view of available resource over time. To make the most of the new low-power infrastructure, it is desirable to design an association scheme not only depending on the SINR, but also the load of BSs. The load may be considerably more balanced with load-aware association, as illustrated in Fig. 1.2b (which uses a max-sum-log-rate wise association).

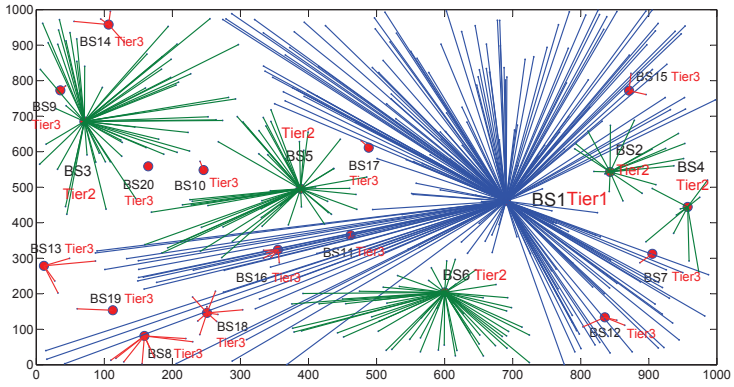
Integration of D2D communication further complicates the network. Potential D2D data can either be transmitted directly (D2D), or via a BS

– termed *mode selection*. Offloading D2D traffic from cellular network to D2D communication may provide a higher spectral efficiency which profits from the short-range transmission, and help alleviate the network congestion, but it either introduces more interference if resources are shared among D2D and cellular links, or consumes some resources that are otherwise available for cellular communication if orthogonal resources are assigned to D2D and cellular links. It is not *a prior* clear when a potential D2D link should transmit directly or be relayed by the BS.

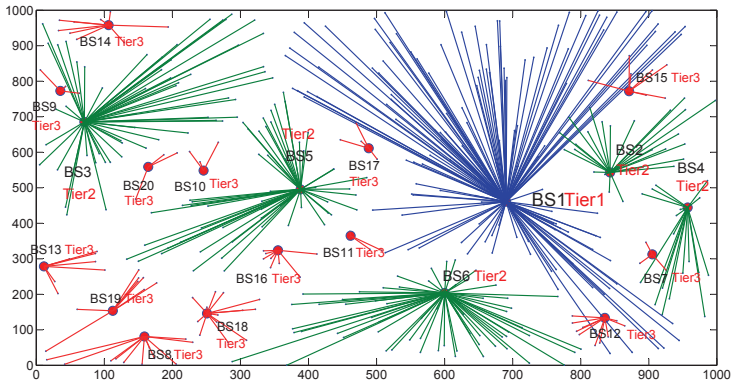
1.2.2 Interference Management

For HetNets with small cells, offloading users from congested macro BSs to lightly-loaded small cells via load-aware association balances the load, but leads to not only weaker signal but also stronger interference at these offloaded users. This implies the importance of joint investigation of interference management along with the load-aware user association. The existing interference management techniques can be broadly classified into two types: uncoordinated and coordinated interference management [16, 17].

Uncoordinated interference management. Fractional frequency reuse (FFR) [18, 19], distributed power control [20, 21], and static resource allocation are three most popular uncoordinated interference management techniques. As the name implies, the techniques in this category do not require coordination among cells, and thus may reduce the overhead and lighten the backhaul burden. On the other hand, the main shortcoming of such simple



(a) Max-SINR association



(b) Load-aware association

Figure 1.2: Max-SINR association versus load-aware association. Lines indicate the user association and highlight the differences in cell association policies.

techniques is the limited performance gain versus more intelligent coordinated techniques [22]. Moreover, techniques like FFR and static resource allocation may require prior careful resource planning, which becomes increasingly difficult as the network with small BSs becomes denser and more irregular.

Coordinated interference management. Different coordinated interference management techniques require coordination of different levels, and thus introduce different amounts of additional overhead. For example, coordinated multi-point transmission (CoMP) as an extension of the multi-user multiple-input and multiple-output (MU-MIMO) technique is proposed as one of the core features for coordinated interference management in LTE-Advanced, which includes different categories of transmission schemes such as dynamic point selection (DPS), dynamic point blanking (DPB), joint transmission (JT) and coordinated scheduling/beamforming (CS/CB) [23–25]. The schemes such as JT with joint precoding and CS/CB introduce excessive overhead and require tight time/frequency synchronization among collaborating cells, while the simpler eICIC technique [26] requires limited overhead and less accuracy in time/frequency synchronization. By choosing low-overhead techniques or minimizing the required overhead, the coordinated interference management is expected to bring greater performance gain than the uncoordinated techniques.

For HetNets with D2D communication, the interference environment depends on how the resources are allocated among the D2D and cellular links. When the D2D and cellular transmissions share the resources, though the

resource utilization is efficient, there is mutual interference among D2D and cellular links. Without interference management, the interference may kill the performance gain by allowing D2D communication.

To summarize, the urgent need for redesigning association schemes and mode selection rules along with interference management motivates the following fundamental questions. For HetNets with small cells: (i) Which users should associate to which BSs? (ii) What is the gap between such optimized user association and a simple association scheme based on biased received signal – referred to as *cell range expansion* (CRE) in 3GPP? (iii) How should the interference be managed and what is the corresponding optimal user association? (iv) What is the performance gain from joint consideration of load balancing and interference management? For HetNets with D2D communication: (i) How should the resource be allocated between D2D and cellular links? (ii) When should a potential D2D link transmit directly versus relaying via the BS? (iii) How should the interference be managed in the case where resources are shared among D2D and cellular links? The goal of this dissertation is to address these questions by developing tractable frameworks for optimization and analysis.

1.3 Contributions and Organization

As indicated above, the emerging wireless networks with the integration of small cells and D2D communication demand the rethinking and redesigning of resource allocation, particularly the user association (mode selec-

tion for D2D communication) and interference management schemes. A major challenge regarding such resource allocation problems is that user association problem is in general a massive combinatorial problem requiring extreme high complexity to solve (e.g., via brute force scheme) for large networks. Moreover, the user association and user scheduling problems are coupled with each other, where the user association determines which users would be scheduled together, while the user scheduling determines the achievable resources per user that further impact the user association. Introducing interference management further complicates the joint problem. This dissertation tackles these key technical challenges, with the main contributions summarized as follows.

User association in HetNets with small cells. In Chapter 2, we propose a utility maximization problem to optimize the user association in a HetNet with small cells, where we adopt the logarithmic utility of users' long-term rate, which is similar to the proportional fairness and achieves a desirable tradeoff between opportunism and fair allocation across users, by saturating the return for providing more rate to users with a high rate. Generally, optimization of user association is combinatorial and thus difficult to solve. On the other hand, implementation in practice may resort to the simple CRE scheme. This paper bridges the gap between these approaches through several physical relaxations of the network-wide optimization problem. Chapter 2 also provides a low-complexity distributed algorithm that converges to a near-optimal solution with a theoretical performance guarantee. The simulation results show that CRE loses little versus the optimal utility, if the bias

factors are chosen carefully. Cell-edge users achieve a large (3x) rate gain by offloading from macro BSs versus the max-SINR association.

User association and resource blanking in HetNets with small cells. The rate distribution greatly benefits from load balancing, by which users are offloaded from congested macro BSs to lightly-loaded small cells, despite the resulting loss in SINR. To further improve the network performance, particularly the rate of cell-edge users, Chapter 3 explores the joint optimization of user association and resource blanking at macro BSs, which manages the inter-tier interference by blanking a fraction of time/frequency resources at macro BSs. By relaxing the constraint of unique association where each user associates to at most one BS, the joint problem becomes a convex optimization, and provides an upper bound on the network utility. Chapter 3 shows both theoretically and through simulation that the optimal solution of the relaxed still results in an association that is mostly unique. The optimal association differs significantly when the macro BSs are on or off; in particular the offloading can be much more aggressive when the resource is left blank by macro BSs. Results show that joint optimization of offloading with blanking is important. The rate gain for cell edge users (the worst 5%) is very large (5x) versus max-SINR association without blanking.

Joint optimization of user association and interference management in massive MIMO HetNets. HetNets with BSs equipped with large-antenna arrays that are able to operate in "massive MIMO" regime are envisioned to play a key role in meeting the exploding wireless traffic demands

[27]. Chapter 4 extends the framework proposed in Chapters 2 and 3 with single-antenna transmission to massive MIMO HetNets, where MIMO techniques not only yield large spectral efficiencies but also provide the option of CoMP for interference management. Chapter 4 studies two interference management techniques: resource blanking and JT with local precoding, which allows multiple BSs to transmit the same signal to users simultaneously with precoding based only on local channel state information (CSI). Since massive MIMO instantaneous rates can be predicted *a priori* due to the fact that fast fading is averaged out via the large number of signal paths from the large-antenna arrays at BSs, the user association and scheduling problems are decoupled, allowing us to propose a unified convex utility maximization problem for the joint optimization of user association and resource allocation with both resource blanking and JT. An efficient algorithm providing near-optimal solutions is proposed. Chapter 4 further proposes a simple scheduling scheme to get approximate but implementable results. Simulations reveal that the proposed methods can significantly outperform the optimal user association without interference management, especially at the cell edge (2x rate gain).

D2D-enabled cellular networks using time-frequency hopping. Besides small cells, the D2D communication is also an emerging offloading approach for cellular networks. Chapter 5 develops a flexible and accurate framework for HetNets consisted of D2D communication and conventional cellular networks. As a fundamental issue in the D2D design, the following two spectrum sharing methods are studied in the downlink (DL) system: dedicated

resource allocation where D2D and cellular links use orthogonal resources, and shared resource allocation where D2D links reuse cellular resources. A time-frequency hopping scheme is proposed to schedule D2D links, which randomizes the interference and provides an Aloha-type mode selection. Chapter 5 provides analytical expressions for several important metrics including the coverage probability and the throughput, which are in closed or simple forms, allowing us to easily explore the impact of key parameters (e.g., the load and hopping probabilities) on the network throughput without extensive simulations. The results show that with an optimal spectrum partition between D2D and cellular links, the dedicated network provides larger throughput in the DL than the shared network where BSs cause strong interference to D2D transmissions. The optimal D2D frequency hopping probability to maximize the throughput depends on users' service demands (i.e. the traffic arrival rate): D2D links with more traffic to transmit should be more aggressive in their spectrum access, despite the interference that this generates to the rest of the network. As for the optimal time hopping in considered interference limited heavily loaded networks, all potential D2D links should operate in D2D mode to maximize the total rate. The obtained throughput can be viewed as an optimized lower bound to other more sophisticated scheduling schemes.

Distributed resource allocation for D2D in cellular networks. As a parallel work to Chapter 5, [28] investigates D2D communication in the context of uplink (UL) cellular networks, which shows that the dedicated and shared resource allocation have comparable throughput. As the UL resources

are often under-utilized compared to DL [29], Chapter 6 studies the case where D2D links reuse cellular UL resources, in the effort to improve the resource utilization. In such cases, the success of co-existence of D2D and cellular transmissions depends heavily on the interference management. Chapter 6 proposes to manage the side-effect of introducing D2D communication via dynamic D2D resource allocation, with the objective to maximize the total throughput. Showing such a maximization problem is non-convex and hence intractable, Chapter 6 proposes a distributed approach based on the pricing mechanism, where the BS adapt a signal to the interference from D2D links that is then transmitted to D2D users, who then play a best response (i.e., selfishly) to this signal. The proposed algorithm is computationally extremely efficient, and requires minimal coordination and cooperation among the nodes. Simulation results show that the proposed algorithm converges quickly with low overhead, and achieves a significant throughput gain (about 5x with 10 D2D links per cell and average D2D link length 80m in our simulation setup) versus the cellular networks without D2D links, while maintaining the quality of cellular links at a predefined service level.

Finally, Chapter 7 concludes this dissertation with a summary of key contributions and a discussion of future research directions.

Chapter 2

Load Balancing in HetNets with Small BSs[1]

As indicated in the previous chapter, in HetNets with different types of BSs, even with a targeted deployment where these small BSs are placed in hotspots, most users still receive the strongest signal from the tower-mounted macro BSs. By actively "pushing" some users onto small BSs that are often lightly loaded, not only these offloaded users but also the remaining macro users can get more resources and thus may improve the rate distribution, despite the decreasing SINR at the offloaded users. This chapter investigates optimal and near-optimal solutions of load-aware user association problem, particularly those with simple requirements for coordination and side information. In the remainder of this chapter, HetNets particularly refer to the networks consisted of conventional macrocellular networks and small cells, and we leave the investigation of networks integrating D2D communication to Chapters 5 and 6.

[1]This chapter has been published in [30]. Coauthors Dr. Beiyu Rong, Dr. Yudong Chen and Dr. Mazin Al-Shalash have provided technical suggestions and insights to this work. Dr. Constantine Caramanis and Dr. Jeffrey G. Andrews are my supervisors.

2.1 Related Work

Most prior work on load balancing schemes applies to macrocell-only networks. Networks with small cells are much more sensitive to the cell association policy because of the massive disparities in transmit power and cell sizes, which result in very unequal loads in max-SINR association. That is, if users simply associate with the strongest BS, the difference in load in macrocell networks is constrained since the cells all have roughly the same coverage area. But in HetNets, the opposite is true, making the problem considerably more complex, and the potential gains from load-aware associations larger.

The existing work on cell association can be broadly classified into two groups: (i) Strategies based on *channel borrowing* from lightly-loaded cells, such as hybrid channel assignment (HCA) [31], channel borrowing without locking (CBWL) [32], load balancing with selective borrowing (LBSB) [33,34], etc; (ii) Strategies based on *traffic transfer* to lightly-loaded cells, such as directed retry [35], mobile-assisted call admission algorithms (MACA) [36], hierarchical macrocell overlay systems [37, 38], cell breathing techniques [39, 40], and biasing methods in HetNets [24]. The approach in this chapter is based on traffic transfer. There have been many efforts in the literature towards traffic transfer strategies in macro-only cellular networks. The so-called "cell breathing" technique [39, 40] dynamically changes (contracts or expands) the coverage area depending on the load situation (over-loaded or under-loaded) of the cells by adjusting the transmit power. Sang *et al.* [41] proposed an integrated framework consisting of MAC-layer cell breathing and load-aware

handover/cell-site selection. Cell breathing aims to balance the load among neighboring macrocells, while in HetNets we additionally need to balance the load among different tiers.

A popular approach in conventional networks, related to the direction we propose, is to achieve load balancing by changing the problem to be a convex optimization. Indeed, there is considerable work investigating different utility functions, such as network-wide proportional fairness (PF) [42], network-wide max-min fairness [43], maximization of network-wide aggregate utility by partial frequency reuse and load balancing [44], and α-optimal user association [45]. We adopt the logarithmic function as the utility function, which is similar to proportional fairness, and achieves a desirable tradeoff between opportunism and fair allocation across users, by saturating the reward for providing more resources to users which already have a high rate.

In HetNets, there are a few recent investigations of the cell association problem. A joint optimization of channel selection, user association and power control in HetNets is considered in [46], aiming to minimize the potential delay, which is related to the sum of the inverse of the per-user SINRs, where the SINR takes into account the load when computing the interference. Corroy *et al.* [47] propose a dynamic cell association to maximize sum rate as well as a heuristic CRE algorithm for load balancing. CRE is an effective method to balance the load among high and low power BSs, which is enabled through cell biasing [24, 48]. It is achieved by performing user association based on the biased measured signal, which leads to better load balancing, but the

improvement of load balancing may not overwhelm the degradation in SINR that certain users suffer. Therefore, how to design the biasing factor is an important open problem.

2.2 Contributions and Organization

In this chapter, we present a load-aware cell association method in DL HetNets, that results in the following main contributions.

First, in Sec. 2.4, we undertake an optimization theoretic approach to the load-balancing problem, where we consider cell association and resource allocation jointly. We decouple the joint optimization problem with a general utility function by relaxing the unique association and allowing users to associate with more than one BS – called *fractional association*. This approach provides an upper bound on achievable network utility which can serve as a benchmark. However, in real system, it is much more difficult to implement multi-BS association than unique association. Therefore, we focus on a logarithmic utility maximization problem for unique association, and show that equal resource allocation is actually optimal when channels are static over a sufficiently large time window. This observation allows the coupled problem to reduce to the cell association problem with equal resource allocation, which along with the fractional association relaxation converts the previously intractable combinatorial problem into a convex optimization problem.

In Sec. 2.5, we exploit the convexity of the problem to develop a distributed algorithm via dual decomposition that converges towards the optimal

solution with a guarantee on the maximum gap from optimality. This provides an efficient and low-overhead algorithm for implementation in HetNets.

In Sec. 2.6, we leverage our provably optimal solutions to ask a basic question: how much of the performance gain can a simple policy based on *a priori* bias factors achieve? Our results show that this simple approach gets surprisingly close to the gains of the load-aware utility maximization. The gains from this approach are shown to be very large for most users in the system, with rate gains ranging from 2-3.5x for the bottom half of users. To put this in context, this is a gain on par with what would otherwise be achieved by a doubling or tripling the amount of spectrum for a given service provider. Cell interior users experience little to no rate gain (or a small loss), but this has little relevance in practice since such users are already well-served.

2.3 System Model

In DL cellular networks, the default association scheme is max-SINR, which indeed maximizes the probability of coverage, i.e., $\mathbb{P}(\text{SINR} > \beta)$, where β is a target SINR (or equivalently minimizes the probability of outage, i.e., $\mathbb{P}(\text{SINR} \leq \beta)$).

The key performance metric is the service *rate*, not *SINR*. The instantaneous rate is of course directly related to SINR (e.g., $\log_2(1 + \text{SINR})$), but the overall served rate is then multiplied by the fraction of resources that user gets. Hence, heavily-loaded cells provide lower rate over time, even if they provide a higher SINR. Load balancing problem is very important in HetNets.

In this chapter, we focus on the DL cell association. UL could likely be considered through a similar approach, but is complicated by the use of UL power control, which changes the interference depending on the association. Here, we assume that all BSs have full buffers and slowly changing (or constant) transmit power, which means that transmit power of BSs is fixed over the association time scale and thus is independent of the specific association.

We consider a DL K-tier HetNet, with each tier models a type of BSs. We denote by $\mathcal{B}$ the set of all BSs, and by $\mathcal{U}$ the set of all users. During the connection period, we denote by r_{ij} the achievable spectral efficiency which is generally a logarithmic function of SINR.

$$r_{ij} = f(\text{SINR}_{ij}) = f(\frac{P_j g_{ij}}{\sum\limits_{k \in \mathcal{B}, k \neq j} P_k g_{ik} + \sigma^2}),$$

where P_j is the transmit power of BS j, g_{ij} denotes the channel gain between user i and BS j, which in general includes path loss, shadowing and antenna gain, and σ^2 denotes the noise power level. The association is assumed to be carried out in a large time scale compared to the change of channels. The SINR for association is averaged over the association time and thus it is a constant regardless of the dynamics of channels (i.e., fast fading is averaged out). As for resource allocation (user scheduling), we assume that resource allocation is carried out well during the channel coherence time, and thus channel can be regarded as static during each resource allocation period. This model is applicable for low mobility environment. We leave the stochastic channel analysis for future work. Note that though this chapter is focus on

single-carrier system, our model can be extended to multi-carrier system in a straightforward manner (i.e., let r_{ij} be the average spectral efficiency over different bands).

Since each BS generally serves more than one user, users in the same cell need to share time and/of frequency resources. The long-term service rate experienced by a user thus depends on the load of the BS and will therefore be only a fraction of the value r_{ij} multiplied by the total available bandwidth in the network (unless BS j exclusively serves user i). We consider a fully loaded system, where the load on a BS is directly proportional to the number of users associated with it.

Moreover, the overall service rate also depends on the resource allocation method of the BSs. In principle, any allocation method or service discipline with which the resource allocation is related to both the load of BSs and the rate of each user can be used. Therefore, the achievable overall rate of user i associated with BS j depends on r_{ij}, r_{qj}, and how BS j distributes its resources among its associated users. We focus on finding an optimal resource allocation and optimal cell associations which maximize the utility. During the connection between the BS j and user i, denoting the fraction of resources BS serves user i by y_{ij}, we can define the overall long term rate as follows.

Definition 2.1. *If user i is associated with BS j, the overall long term rate is*

$$R_{ij} = y_{ij} r_{ij}, \tag{2.1}$$

where $\sum_i y_{ij} = 1,\ \forall j$. *We denote the total overall rate of user* i *by* R_i, *where* $R_i = \sum_j R_{ij}$.

In the following, we investigate a utility maximization problem for the overall rate R_i to find the optimal association and resource allocation.

2.4 Problem Formulation

Taking a utility function perspective, we assume user i obtains utility $U_i(R_i)$ when receiving rate is R_i, where the function $U_i(\cdot)$ is a continuously differentiable, monotonically increasing, and strictly concave utility function [49].

2.4.1 General Utility Maximization: Unique Association

We formulate an optimization problem which involves finding the indicators $\{x_{ij}\}$ corresponding to the association (i.e., $x_{ij} = 1$ when user i is associated with BS j, $x_{ij} = 0$ otherwise) and $\{y_{ij}\}$ corresponding to the resource allocation that maximize the aggregate utility function:

$$\begin{aligned} \max_{x,y} \quad & \sum_{i\in\mathcal{U}} U_i(R_i) = \sum_{i\in\mathcal{U}} U_i(\sum_{j\in\mathcal{B}} y_{ij} r_{ij}) \\ \text{s.t.} \quad & \sum_{j\in\mathcal{B}} x_{ij} = 1, \quad \forall i \in \mathcal{U} \\ & \sum_{i\in\mathcal{U}} y_{ij} \le 1, \quad \forall j \in \mathcal{B} \\ & 0 \le y_{ij} \le x_{ij},\ x_{ij} \in \{0,1\} \quad \forall i \in \mathcal{U}, \forall j \in \mathcal{B}. \end{aligned} \tag{2.2}$$

2.4.2 General Utility Maximization: Allowing Joint Association

The indicator variable x_{ij} enforces unique association, which is combinatorial. Moreover, the cell association has to be considered jointly with resource allocation, because resource allocation depends on the association and user association depends on the achievable resource for each user. Therefore, the resulting problem is difficult to solve. While allowing a user to be served by multiple BSs may require more overhead to implement, and hence perhaps may not be viable in practice, it makes the problem more tractable and provides an upper bound on the network performance. In this section, we make the following assumption:

Assumption 2.1. *We assume that users can be associated with more than one BS at the same time.*

Under this assumption, the constraint $\sum_j x_{ij} = 1$ can be eliminated, and hence there is no need for x_{ij} as additional indicators for cell association. The resource allocation variable $y_{ij} \in [0, 1]$ indicates the association, i.e., user i is associated with BS j when $y_{ij} > 0$, otherwise they are not connected.

Therefore, we focus only on the investigation of how resources should be allocated to different users with different rate r_{ij} so as to maximize the utility, instead of considering in conjunction with cell association.

We formulate the *joint association* problem as follows:

$$\begin{aligned} \max_{y} \quad & \sum_{i\in\mathcal{U}} U_i(\sum_j y_{ij} r_{ij}) \\ \text{s.t.} \quad & \sum_{i\in\mathcal{U}} y_{ij} \leq 1, \quad \forall j \in \mathcal{B} \\ & 0 \leq y_{ij} \leq 1, \forall i \in \mathcal{U}, \forall j \in \mathcal{B}. \end{aligned} \tag{2.3}$$

Note that this joint association scheme focuses on how to allocate resources for each BS, rather than how to associate users. In the following sections, we show that with some specific utility functions (e.g., logarithmic utility) and channel conditions (e.g., static channel over an association period), y_{ij} can be directly found without Assumption 2.1 and thus there is no need to decouple x_{ij} and y_{ij} as in this optimization. However, problem (2.3) provides an ultimate limit on achievable network performance for general utility maximizations. Interestingly, our simulation results show that the bound is quite tight in logarithmic utility maximization.

2.4.3 Logarithmic Utility Formulation

Using linear utility functions for throughput maximization results in a trivial solution, where each BS serves only its strongest user. While throughput-optimal, this is not a satisfactory solution for many reasons. Instead, we seek a utility that naturally achieves load balancing, and some level of fairness among the users. To accomplish this, we use a logarithmic utility function. The resulting objective function with logarithmic utility is

$$U_i(R_i) = \log\left(\sum_j y_{ij} r_{ij}\right).$$

The log utility function is concave, and hence has diminishing returns. This property encourages load balancing. This is consistent with the resource allocation philosophy in real systems, where allocating more resources for a well-served user is considered low priority, whereas providing more resources to users with low rates (e.g., in the linear region of the logarithmic function) is desirable. Thus, logarithmic function as a very common choice of utility function is well known as a objective function striking good balance between network throughput and user fairness [50]. Therefore, in the remainder of this chapter, we use a logarithmic utility function.

2.4.4 Analysis of Optimized Resource Allocation

For general utility functions, we proposed one possible tractable model for the joint cell association and resource allocation problem in Sec. 2.4.2, which allows users to be served by multiple BSs. In practice, this is much more difficult to implement than unique association. Therefore, we consider it as a benchmark in this chapter, providing an upper bound on the network utility. With log utility function and the unique association, the objective function of (2.2) becomes

$$\sum_{j\in\mathcal{B}} \sum_{i\in\{k:x_{kj}=1\}} U_i(y_{ij}r_{ij}).$$

Then, we conduct the resource allocation analysis on a typical BS j and the users associated with that BS. The utility maximization problem for

the users associated with BS j is

$$\begin{aligned} \max_{y} \quad & \sum_{i \in \{k: x_{kj}=1\}} \log(y_{ij} r_{ij}) \\ \text{s.t.} \quad & \sum_{i \in \mathcal{U}} y_{ij} \leq 1, \\ & 0 \leq y_{ij} \leq 1 \quad \forall i \in \mathcal{U}. \end{aligned} \tag{2.4}$$

Definition 2.2. *We define the effective load of BS, denoted by K_j, as the number of users associated with it, i.e., $K_j = \sum_{k \in \mathcal{U}} x_{kj}$, where x_{ij} is the association indicator.*

The optimization (2.4) suggests the following proposition.

Proposition 2.1. *The optimal resource allocation is equal allocation for the cases with static channel over a sufficient large time window (e.g., the association period), i.e., $y_{ij} = 1/K_j$.*

Proof. For the cases with static channel over a sufficient large time window, the objective function of (2.3) is

$$\begin{aligned} & \max_{y} \sum_{i \in \{k: x_{kj}=1\}} \log(y_{ij} r_{ij}) \\ & = \sum_{i \in \{k: x_{kj}=1\}} \log(r_{ij}) + \log(y_{ij}), \end{aligned} \tag{2.5}$$

where $\sum_i \log(r_{ij})$ is constant relative to SINR_{ij}. This resource allocation problem is essentially the proportional fair scheduling (i.e., to maximize the log utility in terms of long-term average throughput). Paper [51] shows that the optimal resource allocation is to equally allocate resources to users. For completeness, we give the proof as follows.

The objective function is equivalent to maximize the geometric mean:

$$\max_{y} \sum_{i \in \{k: x_{kj}=1\}} \log(y_{ij}) \Leftrightarrow \max_{y} \frac{1}{N_u} \log\left(\prod_{i}^{N_u} y_{ij}\right) \Leftrightarrow \max_{y} \sqrt[n]{y_{1j} y_{2j} \cdots y_{N_u j}},$$

where N_U denotes the number of users associated with BS j. As the geometric mean is no greater than the arithmetic mean, we have

$$\sqrt[n]{y_{1j} y_{2j} \cdots y_{N_u j}} \leq \frac{y_{1j} + y_{2j} + \cdots + y_{N_u j}}{N_u}, \tag{2.6}$$

where the equality holds if and only if $y_{1j} = y_{2j} = \cdots = y_{N_u j}$. □

Note that in more dynamic settings which take into account time-varying channels over association period, the equal resource allocation may not be optimal. In such cases, we use equal resource allocation (e.g., the result of round-robin scheduling) as the suboptimal scheme and leave the joint optimization of user association and scheduling for future work. The joint optimization of user association and scheduling in massive MIMO scenarios is studied in Chapter 4.

Given the equal resource allocation, the long-term rate for user i is

$$R_{ij} = \frac{r_{ij}}{K_j}, \tag{2.7}$$

so we can rewrite the optimization problem (2.2) to

$$\begin{aligned} \max_{x} \quad & \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{B}} x_{ij} \log\left(\frac{r_{ij}}{\sum_k x_{kj}}\right) \\ \text{s.t.} \quad & \sum_{j \in \mathcal{B}} x_{ij} = 1, \quad \forall i \in \mathcal{U}, \\ & x_{ij} \in \{0, 1\}, \quad \forall i \in \mathcal{U}, \text{ and } \forall j \in \mathcal{B}. \end{aligned} \tag{2.8}$$

When the network is small, the optimal user association can be found through a brute force search. As illustrated in Fig. 1.2 in Chapter 2, the load-aware association that maximizes (2.8) moves traffic from congested macrocells to more lightly loaded small cells. Note that admission control carries out a similar task, where new arrival users will be blocked or forced to other lightly loaded BSs when the potential BS is heavily loaded. However, admission control is performed before a connection is established (i.e., only for new users rather than existing users), and thus cannot achieve an optimal association in terms of load balancing.

2.4.5 Relaxation to Fractional User Association

The above problem is combinatorial due to the binary variable x_{ij}. The complexity of the brute force method is $\mathcal{O}(N_B^{N_U})$, where N_B and N_U denote the number of BSs and of users, respectively. The computation is essentially impossible for even a modest-sized cellular network. To overcome this, we again invoke Assumption 2.1 to allow users to be associated to more than one BS, i.e., fractional association. This physical relaxation reduces the complexity which is no longer combinatorial, and upper bounds the special case where each user is associated with just one BS. It is more difficult to implement fractional association than unique association in a practical system, and thus we adopt a rounding method to revert solutions to unique association. Numerical results in Sec. 2.7 show that there is almost no loss after rounding, and thus the upper bound provided by fractional association is quite tight.

With fractional association, the indicators x_{ij} can take on any real value in $[0, 1]$. The following relaxation of (2.8) is convex:

$$\begin{aligned} \max_{x} \quad & \sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{B}} x_{ij} \log\left(\frac{r_{ij}}{\sum_{k\in\mathcal{U}} x_{kj}}\right) \\ \text{s.t.} \quad & \sum_{j\in\mathcal{B}} x_{ij} = 1, \quad \forall i \in \mathcal{U}, \\ & 0 \le x_{ij} \le 1, \quad \forall i \in \mathcal{U}, \text{ and } \forall j \in \mathcal{B}. \end{aligned} \tag{2.9}$$

To directly solve the convex optimization (2.9), global network information is necessary, which requires a centralized controller. In the following section, we propose a distributed algorithm that only needs local network information and relaxes the coordination requirement among BSs.

2.5 The Distributed Algorithm Based on the Dual Subgradient Method

The centralized functionality for solving the convex optimization problem is usually implemented by a server in the core network for macrocells (e.g., Radio Network Controller (RNC) which carries out resource management in UMTS), which only allows slow adaptation at relatively long timescales and requires coordination among different tiers. Additional issues with centralized mechanisms include excessive computational complexity and low reliability, as any crash on the centralized controller operation will disrupt load balancing. In HetNets, it is usually difficult to coordinate macrocellls and femtocells which are deployed by operators and users respectively. Therefore, a low complexity distributed algorithm without coordination is desirable.

In this section, we propose a distributed algorithm via dual decomposition [52]. The dual problem of (2.9) is decoupled into two sub-problems, which can be solved separately on users' side and BSs' side respectively.

2.5.1 Dual Decomposition

The primal formulation in (2.9) can be expressed in an equivalent form by introducing a new set of variables, the load metric $K_j = \sum_i x_{ij}$.

$$\begin{aligned} \max_{x} \quad & \sum_i \sum_j x_{ij} \log\left(r_{ij}\right) - \sum_j K_j \log\left(K_j\right) \\ \text{s.t.} \quad & \sum_j x_{ij} = 1, \quad \forall i \in U \\ & \sum_i x_{ij} = K_j, \quad \forall j \in B \\ & K_j \leq N_U \\ & x_{ij}, K_j \geq 0, \quad \forall i \in U, \text{ and } \forall j \in B, \end{aligned} \tag{2.10}$$

where the redundant constraint $K_j \leq N_U$ is added for the convergence analysis of of the distributed algorithm (see Theorem 2.1).

The only coupling constraint is $\sum_i x_{ij} = K_j$ in problem (2.10). This motivates us to turn to the Lagrangian dual decomposition method whereby a Lagrange multiplier μ is introduced to relax the coupled constraint. The dual problem is thus:

$$\textbf{D:} \quad \min_{\mu} \quad D(\mu) = f_x(\mu) + g_K(\mu), \tag{2.11}$$

where

$$f(\mu) = \begin{cases} \max\limits_{x} & \sum\limits_{i}\sum\limits_{j} x_{ij}\left(\log(r_{ij}) - \mu_j\right) \\ \text{s.t.} & \sum\limits_{j} x_{ij} = 1 \\ & 0 \le x_{ij} \le 1 \end{cases} \tag{2.12}$$

$$g(\mu) = \max_{K \le N_U} \quad \sum_{j} K_j\left(\mu_j - \log(K_j)\right). \tag{2.13}$$

When the optimal value of (2.10) and (2.11) is the same, we say that strong duality holds. Slater's condition is one of the simple constraint qualifications under which strong duality holds. The constraints in (2.10) are all linear equalities and inequalities, and thus the Slater condition reduces to feasibility [53]. Therefore, the primal problem (2.10) can be equivalently solved by the dual problem (2.11). Denoting $x_{ij}(\mu)$ as the maximizer of the first sub-problem (2.12) and $K_j(\mu)$ as the maximizer of the second sub-problem (2.13). There exits a dual optimal μ^* such that $x(\mu^*)$ and $K(\mu^*)$ are the primal optimal. Therefore, given the dual optimal μ^*, we can get the primal optimal solution by solving the decoupled inner maximization problems (2.12) and (2.13) separately without coordination among the users and BSs.

2.5.2 The Distributed Algorithm

The outer problem is solved by the gradient projection method [54], where the Lagrange multiplier μ is updated in the opposite direction to the gradient $\nabla D(\mu)$. Evaluating the gradient of the dual objective function requires us to solve the inner maximization problem, which has been decom-

posed into two sub-problems f and g. These two sub-problems can be solved in a distributed manner. The tth iteration of gradient projection algorithm is given as follows:

Users' Algorithm:

i) Each user measures SINRs from all BSs, and receives the value of μ_j broadcast by each BS at the beginning of the iteration.

ii) User i connects to BS j^* which satisfies

$$j^* = \arg\max_j \left(\log(r_{ij}) - \mu_j(t)\right). \tag{2.14}$$

If there are multiple maximizers, user i chooses any one of them.

BSs' Algorithm:

Each BS updates K_j and μ_j, and announces the new multiplier μ_j to the system.

i) The solution to (2.13) is

$$K_j(t+1) = \min\{N_U, e^{(\mu_j(t)-1)}\}. \tag{2.15}$$

ii) The Lagrange multiplier μ_j is updated by

$$\mu_j(t+1) = \mu_j(t) - \delta(t)\left(K_j(t) - \sum_i x_{ij}(t)\right), \tag{2.16}$$

where $\delta(t) > 0$ is a dynamically chosen stepsize that is discussed in Sec. 2.5.3.

There is a nice interpretation of μ_j. The multiplier μ works as a message between users and BSs in the system. In fact, it can be interpreted as the price of the BSs determined by the load situation, which can be either positive or negative. If we interpret $\sum_i x_{ij}$ as the serving demand for BS j and K_j as the service the BS j can provide, then μ_j is the bridge between demand and supply, and Eq. (2.16) is indeed consistent with the *law of supply and demand*: if the demand $\sum_i x_{ij}$ for BS j exceeds the supply K_j, the price μ_j will go up; otherwise, the price μ_j will decrease. Thus, when the BS j is over-loaded, it will increase its price μ_j and fewer users will associate with it, while other under-loaded BSs will decrease the price so as to attract more users. Moreover, the function of μ_j (2.16) in the distributed algorithm motivates a rate bias scheme, which is discussed in Sec. 2.6.

Given $x_{ij}(\mu)$ and $K_j(\mu)$, the adjustment (2.16) can be made completely distributed among BSs based on only local information. At each iteration, the complexity of the distributed algorithm is $\mathcal{O}(N_B N_U)$. As for the exchanged information, at each iteration each BS broadcasts its μ_j which is a relatively small real number, and each user reports its association request to only one BS which it wants to connect to. The amount of information to be exchanged in the distributed algorithm is $M(N_B + N_U)$, where M is the number of iterations, while in the centralized method it is proportional to $(N_B \times N_U)$. The gradient method converges fast generally, especially with the dynamic stepsize proposed in Sec. 2.5.3, and thus M is a small number (less than 20 in the simulation). Therefore, even with the requirement of multiple message ex-

changes, the distributed algorithm can still be superior for some cases, such as large scale problems. It is applicable as long as the convergence of distributed algorithm is faster than the association period. After iteratively performing the above steps, the algorithm is guaranteed to converge to a near-optimal solution. This is proved in the next subsection.

2.5.3 Step Size and Convergence

Suppose the stepsize dynamically updates according to the rule

$$\delta(t) = \gamma(t)\frac{D(\mu(t)) - D(t)}{\|\partial D(\mu(t))\|^2}, \quad 0 < \underline{\gamma} \leq \gamma(t) \leq \bar{\gamma} < 2, \tag{2.17}$$

where $D(t)$ is an estimate of the optimal value D^* of problem (2.11), $\underline{\gamma}$ and $\bar{\gamma}$ are some scalars [55]. We consider a procedure for updating $D(t)$, whereby $D(t)$ is given by

$$D(t) = \min_{0 \leq \tau \leq t} D(\mu(\tau)) - \varepsilon(t), \tag{2.18}$$

and $\varepsilon(t)$ is updated according to

$$\varepsilon(t+1) = \begin{cases} \rho\varepsilon(t), & \text{if } D(\mu(t+1)) \leq D(\mu(t)), \\ \max\{\beta\varepsilon(t), \varepsilon\}, & \text{otherwise,} \end{cases} \tag{2.19}$$

where ε, β and ρ are fixed positive constants with $\beta < 1$ and $\rho > 1$ [55].

Thus in this procedure, we want to reach to a target level $D(t)$ that is smaller by $\varepsilon(t)$ over the best value achieved. Whenever the target level is achieved, we increase $\varepsilon(t)$ (i.e., $\rho > 1$) or we keep it at the same value (i.e., $\rho = 1$). If the target level is not attained at a given iteration, $\varepsilon(t)$ is reduced

up to a threshold ε, which guarantees that the stepsize $\delta(t)$ (2.17) is bounded away from zero. As a result, we have the following theorem.

Theorem 2.1. *Assume that the stepsize $\delta(t)$ is updated by the dynamic stepsize rule (2.17) with the adjustments (2.18) and (2.19). If $D^* > -\infty$ where D^* denotes the optimal value, then*

$$\inf_t D(\mu(t)) \leq D^* + \varepsilon. \tag{2.20}$$

Proof. The derivative of function $D(\mu)$ (2.11) is given by

$$\frac{\partial D}{\partial \mu_j}(\mu) = K_j(\mu) - \sum_i x_{ij}(\mu). \tag{2.21}$$

In our primal problem, $K_j = \sum_i x_{ij} \leq N_U$ where N_U is the total number of users. According to (2.21), when K_j and $\sum_i x_{ij}$ are bounded, the subgradient of dual objective function ∂D is also bounded:

$$\sup_t \{\|\partial D(\mu(t))\|\} \leq c, \tag{2.22}$$

where c is some scalar. Thus, our problem satisfies the necessary conditions of Proposition 6.3.6 in [55], which completes the proof. □

2.6 Cell Range Expansion (Biasing)

The proposed approaches above is sensitive to the deployment of users and BSs, i.e., the algorithms have to run again and again in order to keep tracking of changes in networks. In this section, we investigate a simple CRE

which less sensitive to the change of deployments. CRE is proposed a practical way to balance loads in HetNets, since it allows for a simple uncoordinated decision based only on the received power from a given BS [24,56]. It is implemented by assigning a multiplicative SINR bias to each tier of BSs (depending primarily on their transmit power). For example, if a picocell has a 10 dB SINR bias vs. the macrocell BS, a user would associate with it until the SINR delivered by the macro BS is a full 10 dB higher than the picocell. This can be performed by measuring the pilot signals from the BSs within radio range and then simply associating with the one that has the highest *biased* received power. In this section, we investigate whether this simple approach is compatible with optimal performance by solving (2.9).

There are some recent studies on the SINR bias [48, 57], but have not given any theoretical guidance on the "best" biasing factors in the sense of load balancing and/or achieving some optimization criteria. In this section, we evaluate the CRE with SINR bias provided by our optimal user association scheme. Moreover, the distributed algorithm inspires a *rate* bias scheme where the biasing factor is multiplied with the rate instead of SINR. The best SINR biasing factor is obtained by a brute force search based on the optimal FUA, and the best rate biasing factor is derived directly from the optimal μ_j^* in the dual distributed algorithm. The network-wide performance with either biasing factor gets pretty close to the optimal FUA, among which the rate bias performs better than the SINR bias. A more interesting observation is that the biasing factors are *insensitive to the location of BSs and users, which*

makes the bias schemes simple and robust to implement in practice.

2.6.1 SINR Bias

We first consider the SINR bias, where users are associated with the BS which provides the highest *biased* SINR.

Definition 2.3. *Given the biasing factor A_j for BS j, we define the biased SINR received by user i from BS j as*

$$SINR'_{ij} = A_j \cdot SINR_{ij} = \frac{A_j \cdot P_j g_{ij}}{\sum\limits_{k\in\mathcal{B},k\neq j} P_k g_{ik} + N_0}. \tag{2.23}$$

We adopt an identical biasing factor for all BSs in the same tier [24, 48, 57]. Note that setting the biasing factors at all tiers to 1 reduces to the conventional max-SINR cell association, and setting them to $A_j = 1/P_j$ associates users to the BS with the lowest path loss. Biasing under-loaded small BSs, the cells extend the coverage and attract more users, thus resulting in a more fair distribution of traffic. From the simulation results given in Sec. 2.7, we observe that the biasing factors are quite stable as the BS densities change, and the performance of the SINR bias is very close to the optimal FUA.

2.6.2 Rate Bias

According to our load aware association schemes, the best SINR biasing factors are obtained by a brute force search with high complexity. The solution (2.14) of the dual distributed algorithm motivates the more tractable idea of rate bias. According to (2.14), user i is associated with BS j^*, where

$j^* = \arg\max_j(r_{ij}e^{-\mu_j^*})$. Therefore, by setting the rate biasing factor to $B_j = e^{-\mu_j^*}$, the association would be exactly same as the association obtained by the distributed algorithm.

Definition 2.4. *We define the biased rate of user i from BS j as*

$$r'_{ij} = r_{ij} \cdot B_j. \tag{2.24}$$

Through CRE, users are associated with the BS that serves the maximum biased rate r'_{ij}. In rate bias, the biasing factor is in the exponential term of SINR (i.e., $(1+\text{SINR}_{ij})^{B_j}$), which is different from SINR bias where the biasing factor is multiplied directly to SINR (i.e., $A_j\text{SINR}_{ij}$).

In the distributed algorithm, the price variables are different from BS to BS, even for those belonging to the same tier. However, in the investigation of range expansion, just as for SINR bias, we use the same biasing factor for all BSs in a given tier, which is the mean of the optimal multiplier, i.e., $B_j = E[e^{-\mu_l^*}]$, where $l \in j$th tier. The results of rate bias shown in the next section is very close to the optimal solution of (2.9).

2.7 Performance Evaluation

We consider a three-tier HetNet with transmit power $\{P_1, P_2, P_3\} = \{46, 35, 20\}$dBm. The theoretical analysis throughout this chapter is independent of the spatial distribution of BSs. For the simulation, we model the locations of the macro BSs to be fixed, and the locations of the small BSs to

be uniformly and independently distributed in space. This corresponds to operator deployed macros/picocells, and customer-placed femtocells. We assume the location processes across different tiers are independent, with deployed densities $\{\lambda_2, \lambda_3\} = \{5, 20\}$ per macrocell. In modelling the propagation environment, we use a path loss exponent 3 and 4 for macros/picocells and femtocells respectively. We assume lognormal shadowing with a standard deviation $\sigma_s = 8$dB. At room temperature and bandwidth 10MHz, the thermal noise power is $\sigma^2 = kTB = -104$dBm. We then assume that during the connection period between user i and BS j, the user achieves the Shannon capacity rate, i.e., $r_{ij} = \log_2(1 + \text{SINR}_{ij})$.

2.7.1 Loads among different BSs

Fig. 2.1 compares the load distribution with different association schemes. The max-SINR association results in very unbalanced loads: the macro BSs are over-loaded, while small BSs serve far fewer users, with some even being idle. In the fractional association scheme, the load is shifted to the less congested small BSs, which suggests that our objective alleviates the asymmetric load problem. The results after rounding are almost the same as the global optimum obtained by fractional association, showing the effectiveness of the rounding scheme. This occurs because there are few users associated with more than one BSs: most users are not fractional (i.e., associated to more than one BSs). Moreover, the fractional users usually have a strong preference towards one of the BSs. The proposed distributed algorithm and the CRE

scheme also provide near-optimal load distributions.

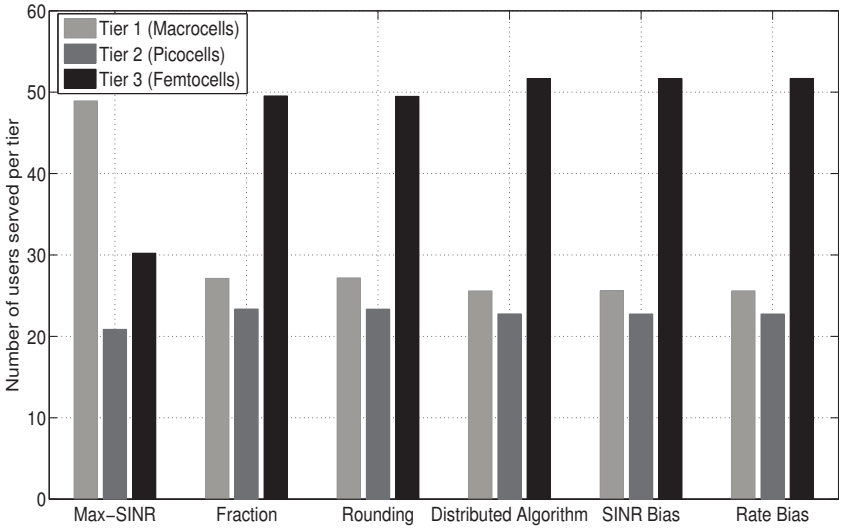


Figure 2.1: Comparisons of average number of users per tier in a three-tier HetNet.

2.7.2 Rate CDF

As another performance measure, Figs. 2.2 and 2.3 show the cumulative distribution function (CDF) of long-term rate in HetNets and conventional networks with different association schemes, respectively. In HetNets, the rate CDFs of joint association, fraction-rounding, the distributed algorithm and CRE all improve significantly (2-3.5x rate gain) at low rate vs. max-SINR association, in both static setting and stochastic setting. The CDFs of fractional-rounding and the distributed algorithm almost overlap, which verifies that the distributed algorithm converges to a near-optimal solution. The result of joint association is very close to the result of fraction-rounding,

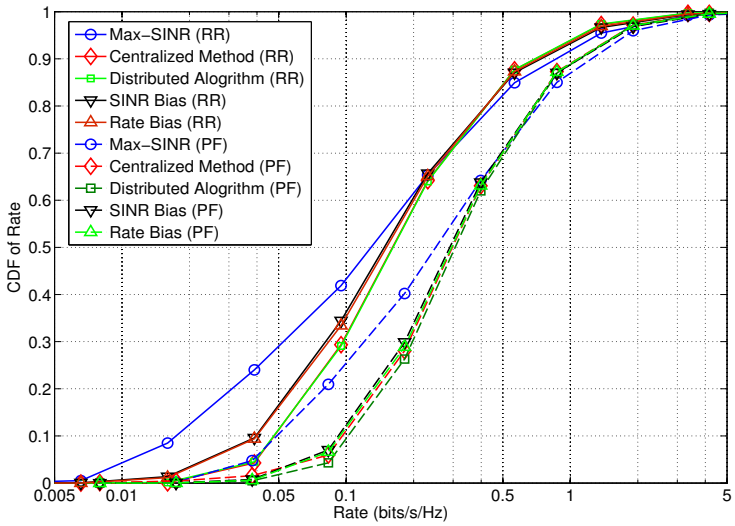


Figure 2.2: The CDFs of overall rate in a three-tier HetNets, in both static setting and stochastic setting. The biasing factors of macro BSs, picos and femtos are $\{A_1, A_2, A_3\} = \{1.00, 4.00, 11.9\}$ in SINR bias, and $\{B_1, B_2, B_3\} = \{1.00, 1.59, 1.88\}$ in rate bias, respectively.

which verifies the conclusions in Prop. 2.1. Note that in stochastic setting, we adopt PF as the scheduling scheme. The static channel equals the average of stochastic channel. From Fig. 2.2, we can see that the rate in stochastic setting with PF is larger than the rate in static setting, although by PF, the resource allocation will eventually converge to almost equal allocation for each user. This is because the channel distribution would be changed by PF (users are more possible to be served in good channel status, i.e., the r_{ij} would be larger than the average rate defined in this chapter). Fig. 2.3 shows that the rate gain is unique for HetNets as long as the users are uniformly distributed.

The ratios of rate α vs. probability $\mathbb{P}(R < \alpha)$ of various approaches

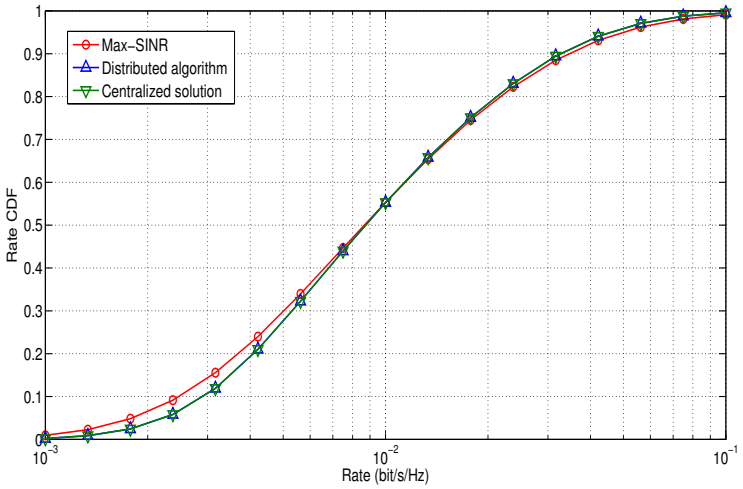


Figure 2.3: The CDFs of overall rate in macro-only networks.

to the max-SINR association are represented in Fig. 2.4. The rate gain is quite large (e.g., 3.5x vs. max-SINR association at the 10% rate point). The results for simple CRE are very close to the optimum associations, where the empirically observed biasing factors of macrocells, picocells and femtocells are $\{A_1, A_2, A_3\} = \{0, 6, 10.8\}$ dB in the SINR bias, and $\{B_1, B_2, B_3\} = \{1.00, 1.59, 1.88\}$ (linear units) in the rate bias, respectively.

2.7.3 Biasing Factor

The effect of BS density and transmit power on biasing factors is considered in Figs. 2.5 and 2.6, respectively. The biasing factors have been normalized, which means that biasing factors at macro BSs are 1.

When the deployed density of small BSs changes, it is interesting to observe in Fig. 2.5 that deploying more small BSs has very little effect on

the biasing factor. Intuitively, though the density of BSs increases, within a reasonable change range of density, there are more users associated with that type of BSs in the optimal association, which makes the needed range expansion almost the same as that in the original scenario. Therefore, *the optimal biasing factors will be almost the same as the network infrastructure deployment evolves.*

However, the story is quite different when the transmit power changes. As the power of 2nd-tier BSs increase in Fig 2.6a, the biasing factor of 2nd-tier BSs steadily decreases, while the biasing factor of 3rd-tier BSs almost stays the same. A similar conclusion can be observed in Fig. 2.6b, where the biasing factor of 3rd-tier BSs decreases gradually and the biasing factor of 2nd-tier BSs is almost static. The biasing factor is smaller as the transmit power increases

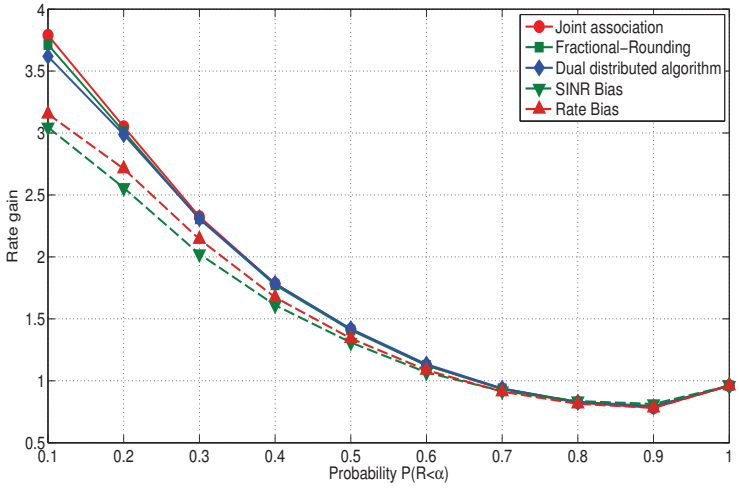
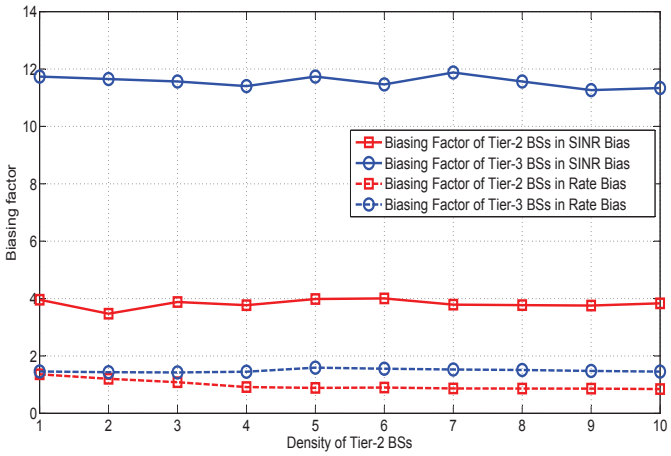
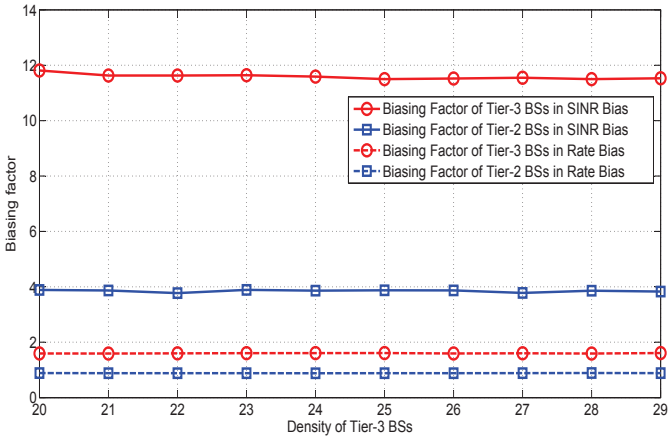


Figure 2.4: Rate gain in a three-tier HetNet. The rate ratio of joint association scheme, fractional-rounding scheme, dual distributed algorithm, and CRE to max-SINR association is represented.



(a) Biasing factors vs. density of tier-2 BSs



(b) Biasing factors vs. density of tier-3 BSs

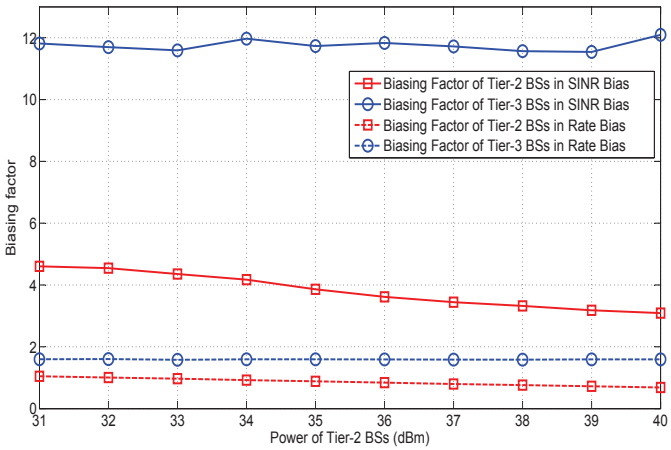
Figure 2.5: Biasing factors vs. density of small BSs in a three-tier HetNet. The density of one tier changes while the others are fixed, with tier 1 always having biasing factor 1.

because users are more likely to be associated to these BSs using max-biased SINR even without a strong bias.

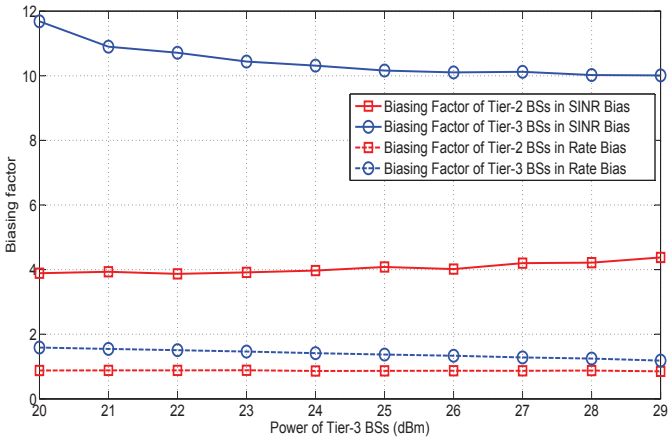
2.8 Summary

In this chapter, we propose a class of novel user association schemes that achieve load balancing in HetNets through a network utility maximization problem. We first consider the cell association and resource allocation jointly, and propose an upper bound on performance. Then we formulate a logarithmic utility maximization problem where the equal resource allocation is optimal for cases with static channels, and design a distributed algorithm via dual decomposition, from the relaxation of physical constraints. The distributed algorithm is proved to converge to a near-optimal solution, with low complexity that is linear to the number of users and the number of BSs. Finally, our scheme is extended to the CRE, which requires limited changes to the existing system architecture by introducing biasing factors to small BSs. We consider two types of biasing factors (SINR and rate), and evaluate the effects of BSs' density and transmit power on the biasing factors in the simulation.

A key observation is that the optimal biasing factors are nearly independent of BS densities for the various tiers, but highly dependent on the per-tier transmit powers. With these optimal biasing factors, the network nearly achieves the optimal load-aware performance. The numerical results demonstrate that a load-aware association significantly improves resource utilization and mitigates the congestion of macro BSs, resulting in a multi-fold



(a) Biasing factors vs. transmit power of tier-2 BSs



(b) Biasing factors vs. transmit power of tier-3 BSs

Figure 2.6: Biasing factors vs. transmit power of BSs in a three-tier HetNet, with tier 1 having biasing factor 1.

gain to the overall rate for most users, particularly those with previously low rates. The network utility maximization problem studied in this chapter will be generalized to address the joint optimization of user association and interference management (e.g., resource blanking) in the following chapters.

Chapter 3

Joint Optimization of User Association and Resource Blanking[1]

As demonstrated in Chapter 2, load-aware user association provides a significant improvement on rate distribution versus max-SINR association. However, load balancing results in the offloaded users experiencing not only a weaker received signal, but also stronger interference. This motivates the straightforward idea of leaving certain time/frequency resources of the macrocells blank (i.e., similar to eICIC, except that eICIC only considers blanking in time domain), during which the offloaded users can receive much higher SINR from the small cells. Although RB blanking decreases the time-frequency resources available to the remaining macrocell users, if there is enough parallelism in the shorter range small cell transmissions (which now have higher SINR and thus rate), this loss can be overcome, and indeed the net gain can be fairly significant [26]. In this chapter, we formulate a joint optimization of user association and RB blanking to address the following questions. How much fraction of resources should the BSs leave blank? How should users as-

[1]This chapter has been published in [58]. Coauthor Dr. Mazin Al-Shalash has provided technical suggestions and many insights to this work, and Dr. Constantine Caramanis and Dr. Jeffrey G. Andrews are my supervisors.

sociate during each of the "on" and "off" periods as a function of the small cell density and other system parameters? What is the best-case gain of such an approach? Similar to Chapter 2, HetNets in this chapter refer to networks with different kinds of BSs (i.e., no D2D communication in this chapter).

3.1 Related Work

There has been increasing effort to study RB blanking in HetNets recently. Paper [59] presents system simulations to compare the performance of RB blanking and CoMP techniques. Some heuristic algorithms for user association and resource partitioning are proposed in [60–62], which improve the cell-edge performance. Papers [63, 64] study the muting ratio leveraging techniques from game theory. Stochastic geometry is also a viable approach to study CRE and RB blanking by providing analytical spatial average results [48, 65, 66]. Besides, a dynamic scheme to update bias factors and RB blanking duty cycle based on reinforcement learning technique is proposed in [67]. The approach related to what we propose is to leverage techniques from optimization theory [30, 68–71]. Paper [68] studies the optimal muting ratio in a simplified scenario with one macro BS and one femto BS, while [69] investigates the optimal RB blanking for dynamic scenarios (e.g., with load variation). Despite so many existing works on RB blanking, the network-wide joint optimization of user association and RB blanking is still far from being fully understood. There are two recent papers [70, 71] focusing on such joint optimization problems which are similar to the problem considered in this

chapter. Different from [70], we further analyze the performance gap between the case where users can associate to multiple BSs – termed *fractional association*, and the case where each user associates to at most one BS – termed *unique association*. Paper [71] considers the joint optimization of RB blanking and user association, where each user either connects to the best macro or best pico BS, which is different from this chapter that considers the user association over any user-BS pairs.

3.2 Contributions

The joint optimization of load-aware user association and RB blanking is a very challenging problem, due to the coupled relationship between user association, scheduling, and RB blanking. This chapter extends the framework proposed in Chapter 2 to the RB blanking case, where all macro BSs are off for the same RBs. The joint optimization is combinatorial if users can only associate with one BS (i.e., unique association), but if this constraint is relaxed to allow users to associate to different BSs (i.e., fractional association), the resulting problem turns out to be convex. It upper bounds the network utility with the unique association. We prove that the number of users associated with multiple BSs is quite limited, and is in fact smaller than the number of BSs. Therefore, a unique association approximated from the fractional association is expected to have comparable performance.

We then turn our attention to the optimal user association during the two different phases – the blank (Off) RBs and normal (On) RBs. We demon-

strate that the optimal associations for blank and normal RBs are very different, with much more aggressive offloading during the Off periods. The fraction of blank RBs is found to increase with the number of picocells in the network. For example, if there are 6-10 picocells per macro, the macrocell should be off about half the time. The gains from joint optimization of load balancing and RB blanking is quite large, while without an appropriately modified association, the gain from RB blanking is limited.

3.3 System Model

Similar to Chapter 2, we consider a downlink HetNet with K tiers of BSs. We consider a synchronous configuration, where each macro BS has the same blank RBs. We jointly optimize the duty cycle of muting at macro BSs and the corresponding user association. The sets of all BSs and users are denoted by $\mathcal{B}$ and $\mathcal{U}$ with size N_B and N_U, respectively. Let $\mathcal{B}_1 \in \mathcal{B}$ be the set of marcocell BSs, with size N_{B_1}. The SINR of user i from BS j in normal (On) RBs is

$$\mathrm{SINR}_{ij}^{(n)} = \frac{P_j h_{ij}}{\sum_{n \in \mathcal{B}/j} P_n h_{in} + \sigma^2}, \quad \forall i \in \mathcal{U}, j \in \mathcal{B}, \tag{3.1}$$

while the SINR of user i from BS j in blank (Off) RBs is

$$\mathrm{SINR}_{ij}^{(b)} = \begin{cases} \dfrac{P_j h_{ij}}{\sum_{n \in \mathcal{B}/(\mathcal{B}_1 \cup j)} P_n h_{in} + \sigma^2}, & \forall j \in \mathcal{B}/\mathcal{B}_1, \\ 0, \quad \forall i \in \mathcal{U}, j \in \mathcal{B}_1, \end{cases} \tag{3.2}$$

where P_j denotes the transmit power of BS j, h_{ij} is the channel gain of the link from BS j to user i, and σ^2 is the noise power level. The channel gain

includes path loss, shadowing and antenna gain. In this chapter, we assume a static channel during each resource allocation period, which is applicable for low mobility environments. Stochastic channel analysis is left as future work.

We denote by $r_{ij}^{(n)}$ and $r_{ij}^{(b)}$ the spectral efficiency of user i from BS j in normal and blank RBs, respectively. Generally, spectral efficiency is a logarithmic function of SINR (e.g., $r_{ij}^{(n)} = \log\left(1 + \text{SINR}_{ij}^{(n)}\right)$). We denote the fraction of resources allocated from BS j to user i in normal and blank RBs by $s_{ij}^{(n)}$ and $s_{ij}^{(b)}$, respectively, where $\sum_{i \in \mathcal{U}} s_{ij}^{(m)} \leq 1, \forall m \in \{b, n\}$. We define the long-term rate as follows.

Definition 3.1. *The long-term rate of user i from BS j is*

$$R_{ij} = (1 - z) s_{ij}^{(n)} r_{ij}^{(n)} + z s_{ij}^{(b)} r_{ij}^{(b)}, \tag{3.3}$$

where z is the fraction of blank RBs. The overall rate of user i, denoted by R_i, can be calculated according to $R_i = \sum_{j \in \mathcal{B}} R_{ij}$.

In the following section, we investigate a utility maximization problem in terms of the long-term rate R_i to find the optimal muting ratio, and the corresponding optimal user association.

3.4 Problem Formulation

The resource allocation variables $s_{ij}^{(n)}$ and $s_{ij}^{(b)}$ also indicate the association (i.e., user i is associated with BS j in normal RBs when $s_{ij}^{(n)} > 0$). Typically, each user will be served by at most one BS, i.e., $\sum_j \mathbf{1}_{\{s_{ij}^{(n)} > 0\}} \leq 1$

and $\sum_j \mathbf{1}_{\{s_{ij}^{(b)}>0\}} \leq 1$. The unique association constraint makes the problem combinatorial, and thus difficult to solve. Though it may not be viable in practice to allow users to be served by multiple BSs at the same time, we relax the unique association constraint and thus make the problem convex, which can serve as an upper bound to benchmark the performance. In the rest of this chapter, we make the following assumption.

Assumption 3.1. *Users can be jointly served by more than one BS at the same time.*

We call users associated with multiple BSs "fractional users". Under the above assumption, the unique association constraint is relaxed, and the resulting optimization problem is:

$$\begin{aligned} \max_{s^{(n)},s^{(b)},z} \quad & \sum_{i\in\mathcal{U}} U_i(R_i) \\ \text{s.t.} \quad & \sum_i s_{ij}^{(n)} \leq 1,\ \forall j, \\ & \sum_i s_{ij}^{(b)} \leq 1,\ \forall j, \\ & s_{ij}^{(n)}, s_{ij}^{(b)} \in [0,1],\ \forall i,j \\ & z \in [0,1],\ \forall j, \end{aligned} \tag{3.4}$$

where $U_i(\cdot)$ is a continuously differentiable, and strictly concave utility function [49]. As revealed in Chapter 2, we adopt a logarithmic utility function, which naturally achieves load balancing. Changing $x_{ij} = (1-z)s_{ij}^{(n)}$ and

$y_{ij} = z s_{ij}^{(b)}$, the optimization problem (3.4) is equivalent to

$$\begin{aligned} \max_{x,y,z} \quad & \sum_{i \in \mathcal{U}} \log \left(\sum_{j \in \mathcal{B}} \left(x_{ij} r_{ij}^{(n)} + y_{ij} r_{ij}^{(b)} \right) \right) \\ \text{s.t.} \quad & \sum_{i \in \mathcal{U}} x_{ij} \leq 1 - z, \ \forall j, \\ & \sum_{i \in \mathcal{U}} y_{ij} \leq z, \ \forall j, \\ & x_{ij}, y_{ij}, z \in [0,1], \ \forall i, j. \end{aligned} \tag{3.5}$$

Proposition 3.1. *The optimization problem (3.5) is convex.*

Proof. Denote the objective function in (3.5) by $g(x, y)$. We will use Hessian matrix to check its convexity. The Hessian has the form

$$\nabla^2 g = - \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & B_{N_U} \end{bmatrix}. \tag{3.6}$$

The matrix B_i can be expressed as

$$B_i = \frac{\mathbf{r}^T \mathbf{r}}{\left(\sum_k \left(x_{ik} r_{1_{ik}} + y_{ik} r_{2_{ik}} \right) \right)^2}, \tag{3.7}$$

where $\mathbf{r} = \left[r_{1_{i1}}, r_{1_{i2}}, \cdots, r_{1_{iN_B}}, r_{2_{i1}}, r_{2_{i2}}, \cdots, r_{2_{iN_B}} \right]$.

Therefore, the matrix B_i is positive semi-definite (PSD) for all i, and thus $-\nabla^2 g$ is also PSD. The problem (3.5) has a concave objective function with linear constraints, which implies that (3.5) is a convex optimization. □

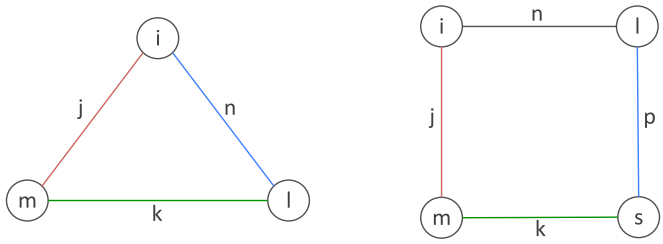
Though the objective function in (3.5) is *strictly* concave with respect to R_i, it is not strictly concave with respect to x and y. We have the following proposition which discusses the uniqueness of the optimal solution.

Proposition 3.2. *The optimization problem (3.5) has a unique optimal resource allocation (i.e., unique x_{ij}^* and y_{ij}^*) almost surely. If $\sum_i \left(x_{ij}^* + y_{ij}^*\right) = 1$, then the optimal solution of problem (3.5) is unique (i.e., z^* is also unique).*

Proof. The proof includes two basic steps. In the first step, $U(R_i)$ is strictly concave in R_i and thus we have unique solution R_i^*. The users with unique association can be obtained uniquely. The second step is to show that the associations of fractional users can also be uniquely generated from R_i^*. This can be proved by bipartite graphs. Details can be found in [72]. When $\sum_i \left(x_{ij}^* + y_{ij}^*\right) = 1$, we have $\sum_{i \in \mathcal{U}} x_{ij}^* = 1 - z^*$ and thus z^* is also unique. □

Returning to Assumption 3.1, what is the impact of the relaxation on the optimal solution? To answer this question, we first use a graph to represent the association, and then by applying *Karush-Kuhn-Tucker* (KKT) conditions, we show that the impact is limited.

In the graph representation of association, the nodes correspond to the users in HetNets, while the edges correspond to the BSs shared between the connected users, illustrated in Fig. 3.1. Each node has a unique ID from 1 to N_U, which is the user index, and each edge has a color from 1 to N_B for BS identification. For example, in Fig. 3.1a, user i is associated with both BSs j and n, and user m is jointly served by BSs j and k. Note that the graph is not necessarily connected. The number of isolated subfigures depends on the number of fractional users. Another important property of the representation graph is that it is comprised of several connected/isolated *complete graphs*.



(a) Example of three users. (b) Example of four users.

Figure 3.1: Examples of graph representation which illustrates the fractional user association.

The convex optimization (3.5) has differentiable objective and constraint functions, and satisfies Slater's condition. Therefore, the KKT conditions provide necessary and sufficient conditions for the optimality [53]. Applying the KKT conditions to problem (3.5), we have the following proposition.

Proposition 3.3. *In the optimal solution, the number of users which are served by multiple BSs in normal RBs is at most* $N_B - 1$. *In blank RBs, the number of users associated with multiple BSs is at most* $N_B - N_{B_1} - 1$.

Proof. We adopt similar techniques in [72]. For completeness, we provide the proof as follows. We define the *Lagrangian* associated with problem (3.5) as

$$\begin{aligned} L(x, y, z, \lambda, \nu) = &- \sum_{i \in \mathcal{U}} \log \left(\sum_{j \in \mathcal{B}} \left(x_{ij} r_{ij}^{(n)} + y_{ij} r_{ij}^{(b)} \right) \right) \\ &+ \sum_{j \in \mathcal{B}} \lambda_j \left(\sum_{i \in \mathcal{U}} x_{ij} - z \right) + \sum_{j \in \mathcal{B}} \nu_j \left(\sum_{i \in \mathcal{U}} y_{ij} - (1 - z) \right), \end{aligned} \tag{3.8}$$

where λ_j and ν_j are the *Lagrange multipliers* associated with the jth inequality constraint in normal and blank RBs in (3.5), respectively. The KKT conditions

are:

$$\begin{cases} \dfrac{r_{ij}^{(n)}}{R_i} = \lambda_j, \text{ if } x_{ij} > 0, \\ \dfrac{r_{ij}^{(b)}}{R_i} = \nu_j, \text{ if } y_{ij} > 0, \\ \sum\limits_j \lambda_j = \sum\limits_j \nu_j, \text{ if } z \in (0,1), \\ \sum\limits_i x_{ij} \le z, \ \lambda_j \left(\sum\limits_i x_{ij} - z \right) = 0, \\ \sum\limits_i y_{ij} \le 1 - z, \ \nu_j \left(\sum\limits_i y_{ij} - (1-z) \right) = 0, \\ x_{ij}, y_{ij}, z \in [0,1], \lambda_j, \nu_j \ge 0 \end{cases} \tag{3.9}$$

We conduct analysis on normal RBs, and the same conclusion can be extended to the blank RBs. From KKT conditions (3.9), for $x_{ij} > 0, x_{in} > 0, x_{mj} > 0$ and $x_{mn} > 0$, we have

$$\frac{r_{ij}^{(n)}}{r_{in}^{(n)}} = \frac{\lambda_j}{\lambda_n} = \frac{r_{mj}^{(n)}}{r_{mn}^{(n)}}, \tag{3.10}$$

which is true with probability 0. Therefore, it is almost sure that any two users can share at most one same BS (i.e., the number of edges between any two nodes in graph is at most 1). Similarly, we consider an example of three users, illustrated in Fig. 3.1a. There are three possible cases:

1. BSs j, n, k are three different BSs: We have

$$\frac{r_{mj}^{(n)}}{r_{mk}^{(n)}} = \frac{\lambda_j}{\lambda_n} \frac{\lambda_n}{\lambda_k} = \frac{r_{ij}^{(n)}}{r_{in}^{(n)}} \frac{r_{ln}^{(n)}}{r_{lk}^{(n)}}, \tag{3.11}$$

which is true with probability 0.

2. $n = k \neq j$:

   The user m is associated with BS j and n, and the user i is also associated with j and n, which contradicts the result in the two-user example.

3. $j = n = k$: It is possible that these three users are all associated with the same BS, where the representation graph becomes a complete graph.

Therefore, a graph representation of three users contains either a loop with the same color or no loop. We can get a similar result for a graph with more than three users (e.g., Fig. 3.1b). In conclusion, the users associated with the same BS constitute a complete graph with edges having the same color. We can generate a new graph, where each complete graph can be considered as a new node. The new graph has no loops and thus it has the maximal number of edges when it is a tree. The number of edges in a tree is one less than the number of nodes in the tree. Therefore, the maximal number of edges connecting different complete graphs is $N_B - 1$. The number of users associated with more than one BSs equals the number of edges in the new graph, which is no more than $N_B - 1$. We can get similar conclusions for the blank RBs. □

Although we relax the unique association constraint, Proposition 3.3 indicates that the relaxed solution would be close to a unique association. This implies the possibility to get a well-approximated near-optimal unique association solution via rounding. From KKT conditions, we also have the

following conclusion about the difference between associations in normal and blank RBs.

Proposition 3.4. *The number of users which get resources from the same BS in normal and blank RBs is at most* $N_B - N_{B_1}$.

Proof. According to KKT conditions (3.9), if user i and m are associated with BS j at both normal and blank RBs, we have

$$\frac{r_{ij}^{(n)}}{r_{ij}^{(b)}} = \frac{\lambda_j}{\nu_j} = \frac{r_{mj}^{(n)}}{r_{mj}^{(b)}},$$

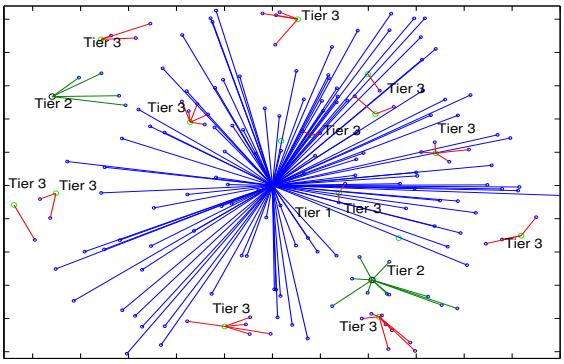
which is true with probability 0. Therefore, it is almost surely true that no more than two users can connect to a BS both in normal and blank RBs. □

Remark 3.1. *Proposition 3.4 implies that the resource allocation in normal RBs is very different from the blank RBs. Only a small fraction of users keep the same association in both normal and blank RBs.*

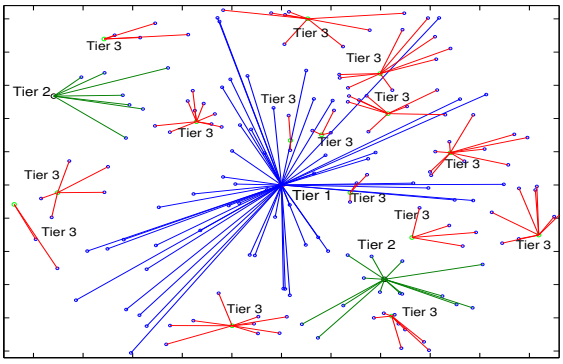
3.5 Performance Evaluation

In this section, we provide simulation results to validate the analytical results and show the rate gain by adopting RB blanking. The main simulation parameters used are summarized in Table 3.1 unless otherwise specified.

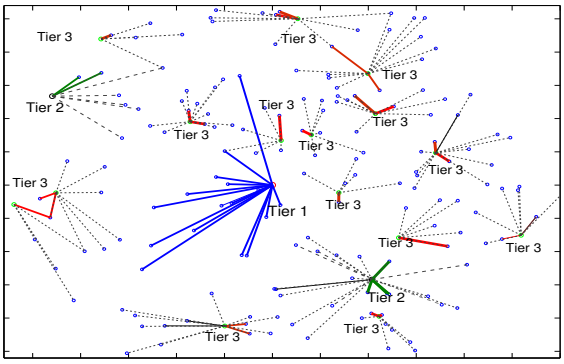
Fig. 3.2 shows examples of associations with different schemes. In the conventional user association scheme shown in Fig. 3.2a, the load is very unbalanced. Most users are associated with the macrocell, but may get small rates even with strong SINRs. The load-aware user association in Fig. 3.2b



(a) Max-SINR association



(b) Load-aware association without RB blanking



(c) Optimal association with RB blanking

Figure 3.2: Examples of associations in HetNets with different association schemes. The dashed lines indicate the association in blank RBs, while the solid lines indicate the association in normal RBs.

Table 3.1: Simulation parameters for user association and RB blanking optimization

Macrocell layout	Hexagonal grid
Pico/femtocell/UE distribution	PPP
Density of macros	$1/500^2$ m^{-2}
Density of picos	$4/500^2$ m^{-2}
Density of femtos	$12/500^2$ m^{-2}
Density of cellular users	$80/500^2$ m^{-2}
Transmit power of macros	40 W
Transmit power of picos	1 W
Transmit power of femtos	0.1 W
Noise power	-104 dBm
Path loss exponent	3.5
Fading	Rayleigh

achieves more balanced load, and thus leads to a more efficient resource utilization. Adopting RB blanking, users can be served in either the normal or/and blank RBs. The associations in blank and normal RBs are illustrated by dashed lines and solid lines in Fig. 3.2c, respectively. We can verify Props. 3.3 and 3.4 that the number of fractional users is very small, and the associations in blank and normal RBs are very different. More users are served by small cells in blank RBs, where there is no strong macrocell interference.

The performance of a three-tier network using different association schemes is compared in Fig. 3.3. We compare five different association approaches, among which the "max-SINR in normal RBs with RB blanking" is a scheme where the association is based on the signal received in normal RBs and the association in blank RBs is kept the same, even though some

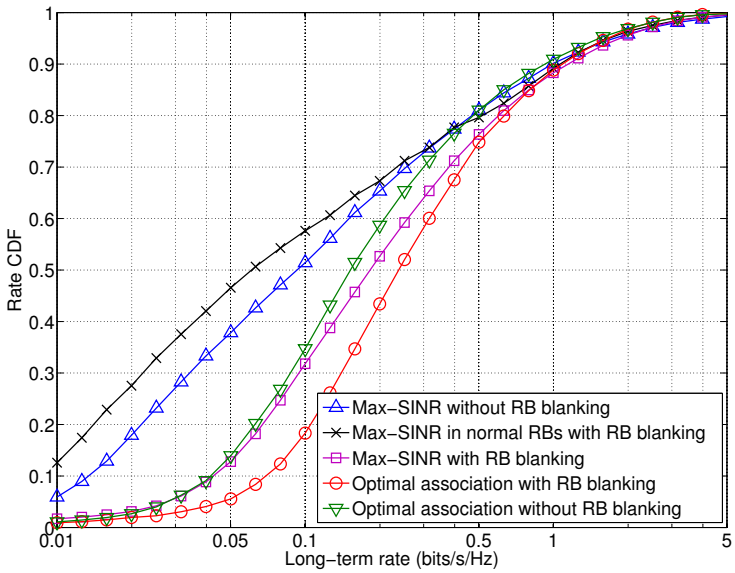


Figure 3.3: The rate distribution of users with different association schemes.

BSs are turned off. By jointly adopting RB blanking, the load-aware association further improves the network performance (e.g., 5x gain for worst 5% users compared to max-SINR without RB blanking). Fig. 3.3 also indicates the importance of appropriate association in networks with RB blanking. By adopting RB blanking with inappropriate association (e.g., max-SINR association based on the received SINR in normal RBs), the network performance may even be degraded. On the other hand, with appropriate association (need not be optimal), the gain can be significant (e.g., 3x gain for worst 5% users using max-SINR with RB blanking compared to max-SINR without RB blanking).

To investigate the impact of different densities of small cells in HetNets, we consider a two-tier network consisting of macrocells and picocells. The fraction of blank RBs in different network settings is shown in Fig. 3.4, where

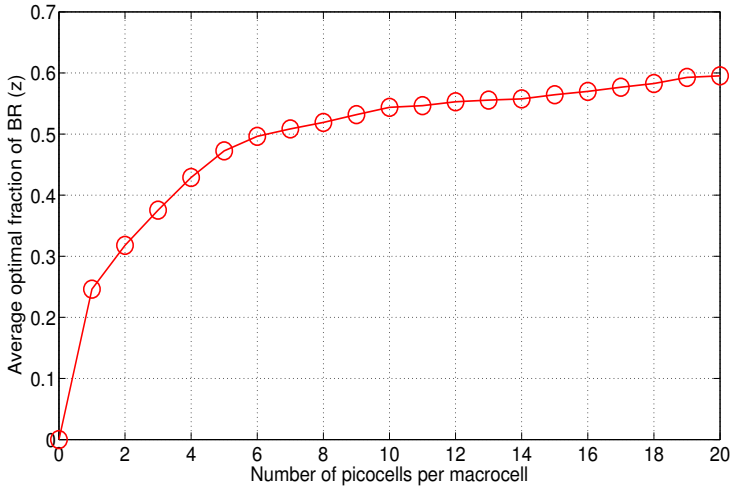
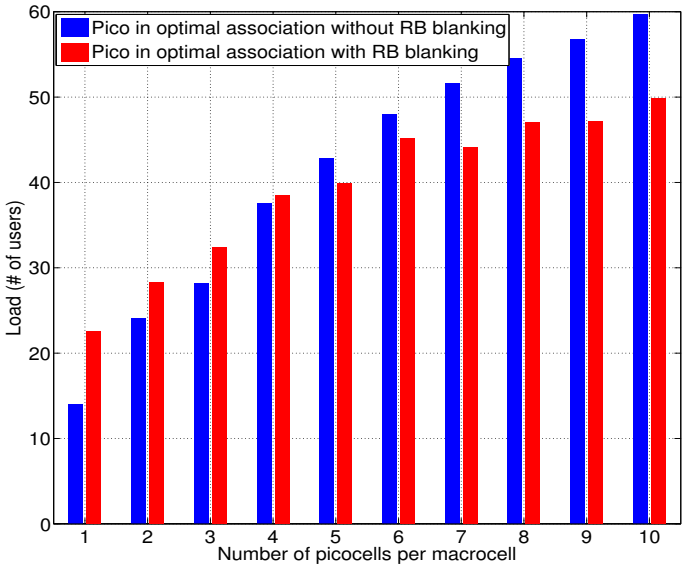


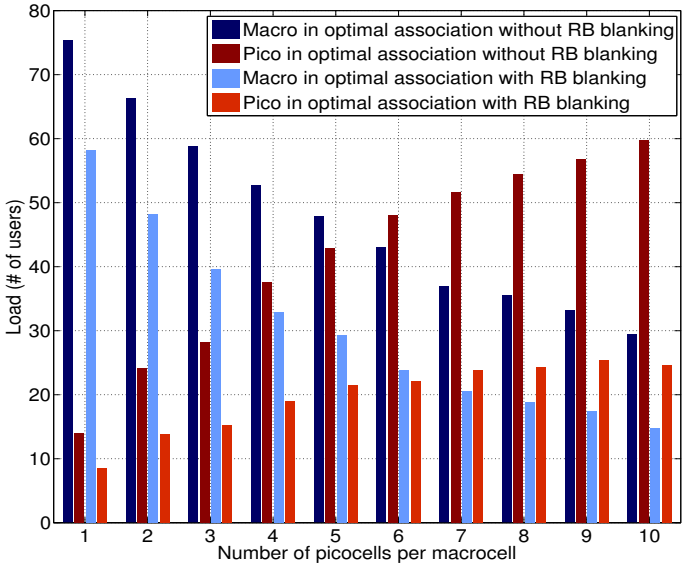
Figure 3.4: The average of optimal fraction of blank RBs (i.e., z) vs. density of small cells.

we average the optimal fraction of blank RBs over different realizations. In Fig. 3.4, the fraction of blank RBs increases as the small cells become denser.

We compare the load in optimal resource allocation with RB blanking to the optimal association without RB blanking in Fig. 3.5, where the load in blank and normal RBs are shown in Figs. 3.5a and 3.5b, respectively. With the increase of picocells, the load at macrocells keeps decreasing and more users are pushed off to small cells, as shown in both figures. Note that adopting RB blanking, a user can be served in blank RBs by picocells and/or in normal RBs by both macrocells and small cells. We have shown that the associations in blank and normal RBs are very different. While more users are pushed off to small cells in both approaches (optimal resource allocation with and without RB blanking) as the density of picocells increases, Fig. 3.5a shows that the picocells serve more users and then fewer users in the blank RBs as the density



(a) Load of optimal association with resource blanking in blank RBs vs. load of optimal association without RB blanking.



(b) Load of optimal association with resource blanking in normal RBs vs. load of optimal association without RB blanking.

Figure 3.5: Load versus small cell density in a two-tier HetNet.

increases, compared to the scenario without RB blanking. On the other hand, Fig. 3.5b shows that RB blanking always decreases the load in normal RBs. One possible reason is that as the density of picocells increases, many users served by small cells in normal RBs already have a good enough rate, so the gain from turning off macro BSs decreases, which decreases the motivation to push off users to picocells in blank RBs. The diminishing gain can also be observed in Fig. 3.6 as the density of picocell increases.

In Fig. 3.6, we show the throughput gain of cell-edge users in different network deployments. The gain is compared to the optimal association without RB blanking (i.e., $\frac{T_a - T_n}{T_n}$, where T_a and T_n are the throughput of worst 10% users using optimal resource allocation with and without RB blanking, respectively). Different from the gain compared to the max-SINR association, the gain here implies the potential impact of RB blanking on the performance improvement. When picocells become denser, it is more necessary to turn off the macrocells, but the gain from RB blanking decreases. In a sparse network, the main interference is from macrocells, and thus the potential gain by turning off macro BSs is large. When the network is increasingly dense, the aggregate interference from small cells keeps increasing and may even overwhelm the interference from macrocells. In this case, though there still exists gain from RB blanking, the SINR improvement of users in small cells decreases due to the large interference from other small cells. Therefore, the gain depends on the aggregate interference from small cells, and thus depends mainly on the transmit power of small cells. Since femtocells have much lower power,

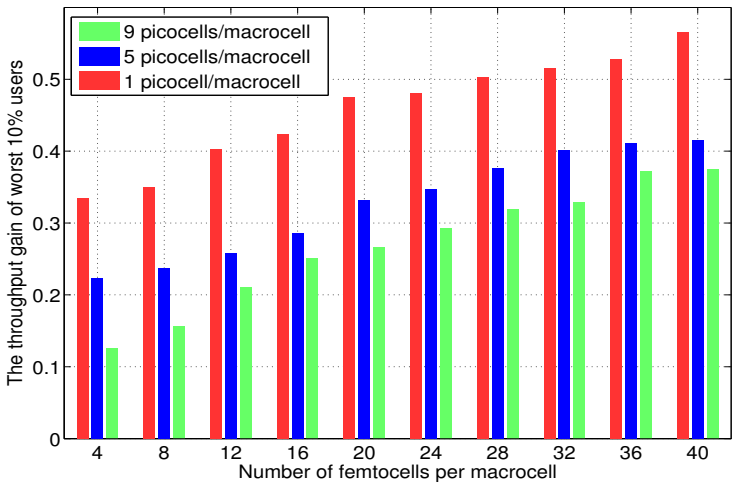


Figure 3.6: The throughput gain of worst 10% users in networks with different deployments of small cells.

the gain keeps increasing as the density of femtocells increases, which is quite different from picocells.

3.6 Summary

In this chapter, we propose a novel framework for the joint optimization of user association and RB blanking in HetNets, which provides a large gain in network performance, in particular "cell-edge" rate. We formulated a network-wide utility maximization problem, which is converted to a convex optimization by unique association relaxation. Although we allow users to be jointly served by multiple BSs, we proved that the number of fractional users is very small (at most $N_B - 1$), and the simulations show that most users have unique association. We also show that associations are very different

when adopting RB blanking. Overall, we show that the load balancing and interference management are key sources of gain in HetNets in Chapters 2 and 3. The developed framework is extended to multiple-antenna transmissions in the following chapter.

Chapter 4

User Association and Interference Management in Massive MIMO HetNets[1]

Two key trends in cellular network evolution are the increasing densification and heterogeneity in BS deployments [74] and the introduction of ever-larger and more densely packed antenna arrays. When the number of antennas is significantly larger than the number of simultaneous users we call this *massive MIMO* [27, 75, 76]. Smart ultra-densification and massive MIMO are considered as two of the most important technologies for 5G cellular systems [77, 78]. This chapter's contribution is to look carefully at this combination[2], particularly in terms of user association, resource allocation, and interference management.

[1] A part of this chapter has been submitted to [73]. Coauthors Dr. Bursalioglu and Dr. Papadopoulos have provided technical suggestions and insights to this work.

[2] As higher-frequency spectrum being available, large arrays become practical even for small cells. For example, at 3.5GHz band, a 36-antenna (arranged on a square grid at half-wavelength separation) can be implemented on a 26cm $\times$ 26cm surface. The required implementation area would be much smaller as the carrier frequency becomes higher.

4.1 Related Work

As discussed in Chapter 1, the max-SINR association results in heavily congested macrocell BSs and lightly loaded low power BSs in HetNets. This results in a very inefficient use of the available time-frequency resources, and strongly motivates load balancing, which in effect means pushing UEs onto lightly loaded small cells even it requires reducing their SINR by many dB [4].

Load Balancing. Several approaches have been used to study load balancing in HetNets, including stochastic geometry [48,79], game theory [80], and system-level simulations [2,24]. Chapter 2 formulated an network utility maximization (NUM) problem for user association in HetNets with single-antenna BSs, where the resources are equally allocated among users in the same cell [30]. Meanwhile, in industry, proactive load balancing is accomplished by *biasing* UEs by some factor (e.g. 10 dB) towards the small cells [2,24]. We also showed in Chapter 2 that this simple suboptimal CRE approach can perform surprisingly close to an optimal association if the right bias values are used. One shortcoming of the work in Chapter 2 is that this equal resource allocation among UEs can be suboptimal if the user associations happen on a much slower time scale than the channel variations. In general, the user association and resource allocation (i.e. scheduling) problems are coupled: the user association determines which users a BS should schedule, while scheduling determines how many resources are available per user, which is a key factor in the user association. Thus, it is very difficult to jointly optimize the user association and resource allocation in general cases.

Interference Management. As the macrocells that the users are offloaded from now become strong interferers, the offloaded users suffer increased interference, which eats into the gains offered by load balancing. This motivates us to jointly consider user association and interference management, to improve the performance of these low-SINR (i.e. "cell edge") users. One popular approach is the resource blanking approach discussed in Chapter 3. Several works have considered the joint problem of user association and RB blanking, for example [69] proposes a dynamic approach adapting the muting duty cycle to load variations while [58, 66, 70, 71, 81, 82] consider a more static approach.

Massive MIMO. A key benefit of massive MIMO is that the extra diversity afforded by the large antenna array averages out the fast fading, and so the instantaneous rate stabilizes to the longer-term mean. This rate is of course still subject to changes in path loss and shadowing, but these happen on much slower time scales, allowing the user association and resource allocation to be decoupled [83]. MIMO techniques also provide the option of serving a user at high rates from multiple BSs – referred to as CoMP, which is proposed as one of the core features in LTE-Advanced [23–25, 84]. The set of BSs that cooperatively serve the same user is called a *BS cluster*. Papers [85, 86] study how to determine the BS clusters, while papers [87–90] investigate the joint design involving either BS cluster selection, beamforming, user scheduling or power allocation.

Cross-layer Optimization. Overall, the joint optimization of user as-

sociation and interference management is still an open issue in massive MIMO HetNets. In this chapter, we jointly consider the user association and both RB blanking and *joint transmission* (JT) as the interference management approaches. The JT is a distributed-MIMO form of CoMP, where each BS's precoding vector depends only on the CSI of its own users and thus does not require CSI exchanges among BSs. Other interference management approaches can also be adopted, but at the cost of overhead, complexity or/and intractability (e.g., JT with joint precoding [25]), and we leave the study of more general interference management approaches (see e.g., [91]) for future work.

To study the joint user association and interference management problem, we propose to use the cross-layer optimization approach, aiming to improve the rate distribution, particularly, the cell-edge performance. Cross-layer optimization is a very popular approach to study the resource allocation problems in wireless networks (see e.g. [92–96] and references therein). Most of the existing literature working on the resource allocation (user scheduling) problems leverage the conclusion given by [1]. That is, the general maximization problem in terms of long-term metric can be decoupled to maximization problems at each slot based on the gradient algorithm [1] – called the *slot-based gradient algorithm* in this chapter. For example, the slot-based gradient algorithm has been used in the cross-layer design of multihop networks [97, 98], CDMA networks [41, 99] and OFDM networks [44, 70, 90, 100–105].

Most existing work studies the cellular transmission, where each user

can be served by at most one BS at any RB. Most existing work studies the cellular transmission, where each user can be served by at most one BS at any RB. Different from the cellular transmission, we aim to design BS clusters in addition to the resource allocation, in order to maximize the network utility, particularly, the rate of cell-edge users. There is one recent related work investigating the BS clustering and user scheduling problems [90], which considers disjoint clusters (i.e., each BS belongs to at most one cluster on any RB) with pre-defined user association. Different from [90], our framework can be applied to not only the cases with disjoint clusters, but also more flexible cases where a BS may belong to different clusters on the same RB, and different users can be served by different clusters (i.e., user-specific clusters). Note that the solution of our formulated problem is always realizable via suitably designed schedulers in the disjoint-cluster cases, but there are two main types of price that have to be paid for this design choice. The first is the potential decrease in the network utility by only allowing disjoint clusters. The second is the complexity of cluster design with the requirement of a careful study on the network topology and shadowing conditions, which becomes increasingly difficult as the network becomes denser and more irregular. Thus, in this chapter, we focus on the user-specific cluster case where BSs can belong to different clusters on the same RBs. To the best of our knowledge, this is the first work proposing a framework to jointly design user-specific clusters and resource allocation to maximize the utility in massive MIMO HetNets.

Leveraging the slot-based gradient algorithm [1], we can transform the

utility maximization problem in terms of long-term metrics to a sequence of maximization problems, with each problem per slot. However, due to that BSs can belong to different clusters and resource constraints cross over different clusters, it is difficult to transform our problem to some typical types of problems that are polynomial-time solvable (e.g., the max weight matching problem [106–111]). Also, note that it is quite difficult to efficiently obtain the achievable optimal utility in [90], let alone characterizing the optimal utility of our more flexible but complicated problems based on the slot-based gradient algorithm. Alternatively, in this chapter, we first investigate the desired average resource allocations by solving a convex maximization problem. Though the obtained resource allocation may not be realizable via a scheduler, it upper bounds the network performance and can serve as a benchmark. We then propose a scheduling scheme that provides approximate but implementable results. The numerical results show that the gap between the performance of the proposed algorithm and the upper bound utility is quite small, which implies that the proposed scheduling scheme can approach near-optimal performance. The theoretical guarantee on the performance gap between the proposed algorithm and the optimal result is beyond the scope of this chapter, and we leave it for future work. The main contributions are summarized in the following section.

4.2 Contributions and Organization

In this chapter, we present a novel framework for the joint optimization of user association and interference management in massive MIMO HetNets, resulting in the following main contributions.

Spectral efficiency analysis. In Sec. 4.4, we derive the instantaneous rate with JT, and the instantaneous rate during normal and blank RBs with RB blanking, for both linear zero-forcing beamforming (LZFBF) and maximum ratio transmission (MRT, also known as conjugate beamforming) in the massive MIMO regime. The fast fading is averaged out, and thus the instantaneous rate is predictable regardless of the scheduling, which is the key difference versus the case without massive MIMO.

A unified network utility maximization (NUM) problem. By exploiting the predictable instantaneous rate, user association and resource allocation problems can be decoupled, allowing us in Sec. 4.5 to formulate a unified convex optimization problem for user association and resource allocation problems with both JT and RB blanking. Note that in the considered JT, the clusters are user-specific (i.e., different users can be served by different clusters). The formulated problem can also be applied to the case where macro and small BSs use orthogonal resources. With blanking but without JT, the optimal solutions can always be realized by a suitably designed scheduler. On the other hand, with JT, we show that there exist some solutions that are not implementable. Naturally, the solution of the NUM problem – called the *NUM solution* – upper bounds the network performance and can serve as a

useful benchmark.

Dual subgradient based algorithm. Sec. 4.6 presents an efficient algorithm based on the dual subgradient method, which converges to near optimal dual variables. Since the objective function is not strictly convex, it is difficult to obtain optimal primal variables given optimal dual variables. Exploring the solution structure, we formulate a small-size linear programming (LP) to get the optimal primal variables. The proposed algorithm can be implemented in a partially distributed manner with low overhead.

Simple scheduling scheme to approach the NUM solution. In Sec. 4.7, we develop a scheduling scheme to yield the resource allocations closely matching the NUM solution, by approximating the NUM solution to the results with unique association (i.e., users are served by at most one cluster on each RB). Showing the limited number of users connecting to multiple clusters per RB in heavily loaded networks, it is expected that the approximate resource allocations with unique association are near the NUM solution.

Simulation results in Sec. 4.9 show a significant gain by jointly optimizing user association with interference management. For example, the rate of bottom (the 10th percentile) users in our setup is about 2.2x with respect to the optimal user association in cellular transmission, which itself is much larger than the max-SINR association. The dual subgradient based algorithm approaches the NUM solutions. Also, the proposed scheduling scheme has near optimal performance, within 90% of the NUM solution.

4.3 System Model

In this chapter, we focus on delay-tolerant best-effort traffic. We consider a DL HetNet with J BSs and K single-antenna users. We let $j \in \mathcal{B} = \{1, 2, \ldots, J\}$ and $k \in \mathcal{U} = \{1, 2, \ldots, K\}$ be the index of BSs and users, respectively. We denote by M_j the number of antennas at BS j with $M_j \gg 1$. We assume time division duplex (TDD) operation with reciprocity-based CSI acquisition [75, 112]. Hence, each user sends a single UL pilot to train all nearby BS antennas. This enables the training of large antenna arrays with overhead proportional to the number of simultaneously served users. In contrast to feedback-based CSI acquisition, it also allows a user to train *multiple nearby* BSs, which enables CoMP without additional training overhead.

4.3.1 Channel Model

We denote the transpose, conjugate and conjugate transpose of matrices by $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^H$, respectively. We denote the set of BSs and of users by $\mathcal{B}$ and $\mathcal{U}$, respectively. The transmit power of BS j is denoted by P_j. We assume a block-fading channel model where the channel coefficients remain constant within each RB [75, 76, 112, 113]. On a generic RB, we denote the channel matrix between BS j and users by $\mathbf{G}_j$, with the kth column being $\mathbf{g}_{kj} = [g_{kj,1}, \cdots g_{kj,M_j}]^T$, where $g_{kj,i} = \sqrt{\beta_{kj}} h_{kj,i}$ is the channel between the ith transmit antenna of BS j and user k, which includes both slow fading β_{kj} and fast fading $h_{kj,i}$. The slow fading β_{kj} characterizes the combined effect of distance-based path loss and location-based shadowing. Vectors $\mathbf{h}_{kj} =$

$[h_{kj,1}, \cdots, h_{kj,M_j}]^T$ capture the fast fading. We assume each link experiences independent Rayleigh fading, i.e., $\mathbf{h}_{kj}$ are complex Gaussian i.i.d. random variables. We let $\mathbf{F}_j$ denote the precoding matrix at BS j, whose kth column $\mathbf{f}_{kj}$ is the beam (i.e., the precoding vector) for user k. The signal symbol of user k is denoted by s_k, where s_k has unit energy. The thermal noise at user k is denoted by w_k, which is assumed to be additive white Gaussian noise (AWGN) with variance σ^2.

4.3.2 Admissible Joint Transmission

With massive MIMO, a subset of users are scheduled for transmission on each RB. In particular, the coded data for any given user can be transmitted either from a single BS – called *cellular transmission*, or from multiple BSs via a CoMP scheme – referred to as JT. The JT poses many challenges making the extension of cellular transmission to JT nontrivial, where key challenges and issues include additional overhead for CSI exchanges among cooperating BSs, users' various preferences to different clusters, and the dependence of a user's spectral efficiency on the other users served by the same BS. To address these challenges, we consider the following particular form of JT that allows harvesting performance gains at the cell edge with low operational overhead, where we denote by S_j the number of users that can be simultaneously served by BS j in cellular transmission.

Definition 4.1. Admissible Transmission Schemes (ATSs): *An ATS schedules users for transmission on a sequence of RBs, and satisfies the fol-*

lowing on each RB:

(1) All users served by a given BS j are served in clusters of the same *size L for some $L \geq 1$;*

(2) BS j in clusters of size L serves at most $S_j(L)$ users with $S_j \leq S_j(L) \leq LS_j$ and $M_j \gg S_j(L)$;

(3) The user beams (i.e., precoding vectors) at BS j are designed as if BS j was in cellular multi-user (MU)-MIMO transmission over all the users it serves;

(4) Each BS serving a user transmits the same coded user stream on a beam that is (independently and locally) designed for users at that BS.

We assume each BS equally allocates its power to the set of scheduled users. The corresponding spectral efficiency can be viewed as a lower bound for cases with power optimization. Table 4.1 provides an ATS example complying with Defn. 4.1, involving clusters of size 1 (cellular transmission) and 2. Four BSs are considered with $P_j = 1$, $S_j(1) = S_j = 2$ and $S_j(2) = 3$. As the table reveals, each BS on RB #1 engages in cellular transmission. On RB #2, pairs of BSs perform JT with each BS pair serving a triplet of users. RBs #3 and #4 provide additional more interesting modes. On RB #3, no user is served by the same cluster on RB #3. On RB #4, BSs 1 and 2 serve users in clusters of size 2 while BSs 3-4 serve users in cellular transmission. Note that if orthogonal pilots are used, (at least) 8, 6, 6 and 7 uplink pilot dimensions

Table 4.1: Example of RBs enabled by admissible transmission schemes over 4 BSs.

RB		BS 1	BS 2	BS 3	BS 4
	Cluster Size	1	1	1	1
#1	User Power	1/2	1/2	1/2	1/2
	Served Users	1,2	3,4	5,6	7,8
	Cluster Size	2	2	2	2
#2	User Power	1/3	1/3	1/3	1/3
	Served Users	1,2,3	1,2,3	4,5,6	4,5,6
	Cluster Size	2	2	2	2
#3	User Power	1/3	1/3	1/3	1/3
	Served Users	1,2,3	1,4,5	2,4,6	3,5,6
	Cluster Size	2	2	1	1
#4	User Power	1/3	1/3	1/2	1/2
	Served Users	1,2,3	1,2,3	4,5	6,7

(one dimension per user) are needed to enable RBs #1, #2, #3 and #4, respectively. Evidently, the choice of scheduled user sizes $S_j(L)$ signifies how aggressively pilot dimensions are reused across the network (e.g., S_j for fully reused pilots and LS_j for orthogonal pilots).

It is worth making a few remarks regarding the choice of ATSs in Defn. 4.1. First, the schemes of Defn. 4.1 provide the following CoMP benefits:

(1) **Performance gains at the cell edge:** The beamforming (BF) gain becomes intra-cluster BF gain in JT, as the same coded data is transmitted from all BSs serving the user. Similarly, the intra-cell interference mitigation is extended across the cluster of BSs by which the user is served. As a result,

performance gains can be realized at the cell edge.

(2) **Low training overhead:** An UL pilot from a user can be received at all nearby BS antennas, whether these are in the same or different locations. Thus, the CSI acquisition between a user and nearby BSs need not incur additional overhead with respect to cellular transmission in TDD systems.

(3) **BSs in JT may serve more users simultaneously than in cellular transmission:** Recall that the service capability (i.e., the number of simultaneously served users) of BS j in clusters of size L, $S_j(L)$, essentially depends on the number of available UL pilot resources. Thus, BSs in clusters may serve more users than in cellular transmission if UL pilots are not fully reused. For example, in Table 4.1 with orthogonal pilots, each BS can serve 2 UEs on RB #1 requiring 8 UL pilots, while each BS can serve 3 UEs on RB #2 requiring only 6 UL pilots. This implies that BSs may serve more users at each RB in JT versus the cellular transmission, but the power allocated from each BS to each user is reduced.

In addition, ATSs possess several important properties that are not present in general CoMP schemes:

(a) **Local precoding at each BS:** This is due to (iii) in Defn. 4.1. For instance, the beam with LZFBF for each user served by BS j is chosen within the null space of the channels of all the other users served by BS j, no matter whether there are additional BSs serving the user on the same RB or not.

(b) **No need for CSI exchanges among BSs:** Due to local precoding, BS

j only needs CSI between the users it serves and the antennas of BS j in order to generate the user beams at BS j. In contrast to general JTs [25, 114, 115] that design beams depending on CSI between *other users* and all BSs in the cluster and thus require global CSI, the proposed JT only requires local CSI and does not introduce additional overhead.

(c) **Flexible scheduling allowing user-specific clusters:** Revoking the local precoding again,each BS only needs to know which subset of users to serve on each RB. This allows users to be served by overlapping but different BS clusters on the same RBs (see, e.g., RB #3 in Table 4.1).

(d) **Predictable instantaneous rates:** As shown in [115], the instantaneous user rates can also be predicted *a priori* with CoMP. However, unlike the general CoMP schemes, where a user's instantaneous rate depends on the other users scheduled on the same RB [115], the instantaneous rate in ATSs is *independent* of the other users in the scheduling set.

4.3.3 Admissible Transmission with blanking

We call the set of clusters that can transmit on the same RB an *admissible transmission mode* (ATM). Thus, the ATSs given by Defn. 4.1 refer to schemes with orthogonal RBs allocated to different ATMs. To incorporate the resource blanking into ATSs given by Defn. 4.1, we define normal and blank RBs as different ATMs. Thus, the set of interfering BSs may be different in different ATMs. We consider a synchronous configuration as in [58], where all macro BSs mute at the same RBs. For example, defining BSs 1 and 2 in Table

4.1 as macro BSs and BSs 3 and 4 as small BSs, there exists an additional RB type corresponding to blank RBs, that only includes BSs 3 and 4.

4.4 Instantaneous Rate and Long-term Rate

Before formulating the NUM problem, we first develop proxy expressions for the instantaneous rates and for the long-term rates (throughputs) with either LZFBF or MRT.

Without loss of generality, we assume that the BS cluster $\mathcal{C}$ serves user k in an ATM $\mathcal{A}$. We consider a scheduling policy on RBs $\{1, 2 \cdots, T\}$ and assume that all the large-scale coefficients stay fixed within this period. Any such scheduling policy can be described in terms of the scheduling sets $\{\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t); \quad \forall \mathcal{C}, \ \forall t \in \{1, 2 \cdots, T\}\}$, where $\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t)$ denotes the set of active users served by cluster $\mathcal{C}$ on RB t. Thus, the received signal at an active user $k \in \mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t)$ in $\mathcal{A}$ on RB t can be expressed by

$$y_{k\mathcal{C}}^{(\mathcal{A})}(t) = \underbrace{\sum_{j\in\mathcal{C}} \sqrt{\frac{P_j}{S_j(|\mathcal{C}|)}} \mathbf{g}_{kj}(t)^H \mathbf{f}_{kj}(t) s_k}_{\text{desired}} + \underbrace{\sum_{j\in\mathcal{C}} \sum_{\substack{u\in\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t) \\ u\neq k}} \sqrt{\frac{P_j}{S_j(|\mathcal{C}|)}} \mathbf{g}_{kj}(t)^H \mathbf{f}_{uj}(t) s_u}_{\text{intra-cluster interference}}$$
$$+ \underbrace{\sum_{l\notin\mathcal{C}} \sum_{u\in\cup_{(\mathcal{C}'\in\mathcal{A}: l\in\mathcal{C}')}\mathcal{U}_{\mathcal{C}'}^{(\mathcal{A})}(t)} \sqrt{\frac{P_l}{S_l(|\mathcal{C}'|)}} \mathbf{g}_{kl}(t)^H \mathbf{f}_{ul}(t) s_u}_{\text{inter-cluster interference}} + \underbrace{w_k}_{\text{noise}}. \tag{4.1}$$

Note that the set of interfering BSs depends on the ATM $\mathcal{A}$. For clusters that are not in $\mathcal{A}$, we have $y_{k\mathcal{C}}^{(\mathcal{A})}(t) = 0$, and thus the instantaneous rate is zero.

4.4.1 Instantaneous rate

We assume that each BS has perfect CSI. With massive MIMO, the instantaneous rate of user k from BS j on RB t, denoted by $r_{kj}(t)$, can be predicted *a priori*. In particular, there exist deterministic quantities $\{r_{kj}\}$ such that $r_{kj}(t) \overset{\text{a.s.}}{\to} r_{kj}$, $\forall k \in \mathcal{U}$ and $\forall j \in \mathcal{B}$ as $M_j, S_j \to \infty$ with fixed $S_j/M_j \geq 0$ [75, 76, 115, 116]. Moreover, this convergence is very fast with respect to the M_j's. We use the deterministic $\{r_{kj}\}$ as proxies for instantaneous rates.

Adopting LZFBF, $\mathbf{F}_j = \mathbf{G}_j \left(\mathbf{G}_j^H \mathbf{G}_j\right)^{-1} \mathbf{A}_j^{1/2}$ is the precoding matrix at BS j, where $\mathbf{A}_j$ is the normalizing coefficients matrix. In this case, the intra-cluster interference is 0.

Proposition 4.1. *The instantaneous rate of user k from cluster $\mathcal{C}$ in ATM $\mathcal{A}$ using LZFBF can be approximated by*

$$r_{k\mathcal{C}}^{(\mathcal{A})} \approx \log_2 \left(1 + \frac{\sum_{j \in \mathcal{C}} \sum_{l \in \mathcal{C}} \sqrt{P_j P_l \beta_{kj} \beta_{kl} b_j(|\mathcal{C}|) b_l(|\mathcal{C}|)}}{\sigma^2 + \sum_{l \notin \mathcal{C}, l \in \mathcal{C}' \in \mathcal{A}} P_l \beta_{kl}} \right), \tag{4.2}$$

where $b_j(|\mathcal{C}|) = \frac{M_j - S_j(|\mathcal{C}|)+1}{S_j(|\mathcal{C}|)}$.

Proof. Please see Appendix 4.11.1. □

With MRT, the precoding matrix at BS j is $\mathbf{F}_j$ with the kth column being $\mathbf{f}_{kj} = \frac{\mathbf{g}_{kj}}{\|\mathbf{g}_{kj}\|}$.

Proposition 4.2. *The approximate instantaneous rate of user k from cluster $\mathcal{C}$ in ATM $\mathcal{A}$ using MRT is*

$$r_{k\mathcal{C}}^{(\mathcal{A})} \approx \log_2 \left(1 + \frac{\sum_{j \in \mathcal{C}} \sum_{l \in \mathcal{C}} \sqrt{\frac{P_j P_l M_j M_l \beta_{kj} \beta_{kl}}{S_j(|\mathcal{C}|) S_l(|\mathcal{C}|)}}}{\sigma^2 + I_{k\mathcal{C}} + \sum_{l \notin \mathcal{C}, l \in \mathcal{C}' \in \mathcal{A}} P_l \beta_{kl}} \right), \tag{4.3}$$

where $I_{k\mathcal{C}} = \sum_{j\in\mathcal{C}} \frac{(S_j(|\mathcal{C}|)-1)}{S_j(|\mathcal{C}|)} P_j\beta_{kj}$ *is the non-zero intra-cluster interference.*

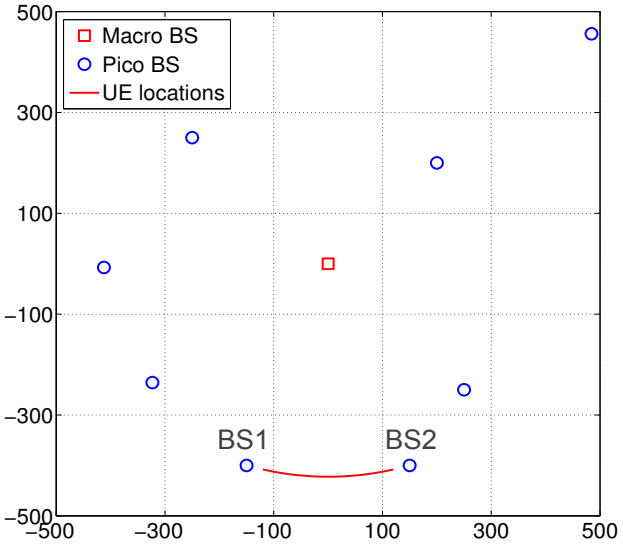
Proof. Please see Appendix 4.11.2,. □

Eqs. (4.2) and (4.3) assume that $\forall j \in \mathcal{C}$, BS j serves $S_j(|\mathcal{C}|)$ users and allocates $P_j/S_j(|\mathcal{C}|)$ fraction of its power to each user. In the case that fewer users are served by one of the BSs, (4.2) and (4.3) represent achievable lower-bound instantaneous rates.

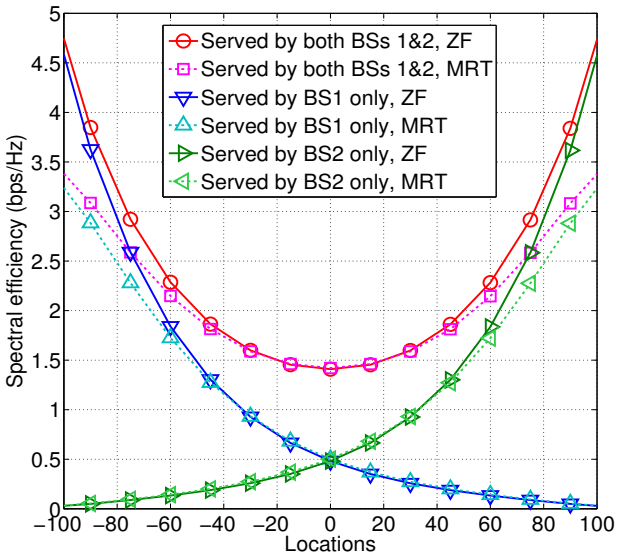
Recalling that macro BSs do not belong to the ATM on blank RBs, the instantaneous rate of macro users on blank RBs is therefore 0, while there is no interference from macro BSs to users in small cells. Obviously, the users associated to small BSs would benefit from larger SINR on blank RBs. On the other hand, for JT we give an illustration example in Fig. 4.1 showing the change of instantaneous rate versus user locations. When the user is close to the origin, which is the cell edge area of BSs 1 and 2, the instantaneous rate from cluster $\{1, 2\}$ is about 3x compared to the case where users are served by an individual BS. This implies the potential benefits of JT for cell-edge users.

4.4.2 Long-term Rate

As discussed in Sec. 4.1, the user association and resource allocation (scheduling) problems is generally coupled with each other. On the other hand, as shown in (4.2) and (4.3), the instantaneous rate in the massive MIMO regime does not depend on the fast fading and user scheduling. Moreover, the UL pilots can be received at all nearby BSs, and thus users can be served by



(a) Illustration of user location



(b) Instantaneous rate vs. x-axis coordinate of the path in (a).

Figure 4.1: Illustration of instantaneous rate versus user locations, when $S_j(|\mathcal{C}|) = |\mathcal{C}|S_j$. The location in Fig. 4.1b indicates the x-axis coordinate of the path in Fig. 4.1a.

different BSs on different RBs [83]. Therefore, there is indeed no conventional "association" concept in massive MIMO HetNets, and users can be fractionally associated with (essentially scheduled by) different BS clusters in the long term. This simplifies the coupled relationship between user association and resource allocation problems, exploiting which we can relax the requirement of *a priori* knowledge of available resources per user given the association (e.g., the equal resource allocation in [30, 79]), and jointly optimize the user association and its corresponding resource allocation as studied in [83].

Let $x_{k\mathcal{C}}^{(\mathcal{A})} = \lim_{T\to\infty} \frac{|\{t:1\leq t\leq T,\ k\in\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t),\ t \text{ is allocated to } \mathcal{A}\}|}{T}$ be the fraction of resources allocated by cluster $\mathcal{C}$ to user k in $\mathcal{A}$ – called *activity fraction*. For the ATSs of interest, we can obtain the long-term rate similar to cellular transmission in [83], which depends on the instantaneous rates and activity fractions from the scheduling policy. In the limit $T \to \infty$, the long-term rate of user k can be expressed as[3]

$$R_k = \sum_{\mathcal{A}} \sum_{\mathcal{C}:\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}\, r_{k\mathcal{C}}^{(\mathcal{A})}. \tag{4.4}$$

4.5 Unified NUM Problem Formulation

4.5.1 Restricting Options of ATMs

Before formulating the problem, it is worth restricting the options of ATMs to obtain solutions that are of practical interest. We focus on clus-

[3]Convergence to the limiting expressions of interest is very quick [83].

ter sizes $L \in \{1, 2, \cdots, L_{\max}\}$ for some appropriately chosen maximum[4] size, $L_{\max}$. Motivated by the example in Table 4.1, we consider the following ATSs.

Definition 4.2. Uniform Cluster-Size Scheme (UCS)*: An ATS from Defn. 4.1 is a UCS if*

(1) $\lambda_{\mathcal{A}}$ fraction of RBs is allocated to ATM $\mathcal{A}$, with $\sum_{\mathcal{A}} \lambda_{\mathcal{A}} \leq 1$;

(2) on any RB in the $\lambda_{\mathcal{A}}$ fraction, the scheduled users are served by (user-dependent) clusters of the same size, denoted by $L_{\mathcal{A}}$;

(3) in ATM $\mathcal{A}$, each BS does not serve more than $S_j(L_{\mathcal{A}})$ users.

In the UCS, users served by clusters of different sizes are scheduled on orthogonal RBs. For the example in Table 4.1, such a scheme enables scheduling policies on RBs of types #1, #2 and #3, but not of type #4. The possible ATMs in the UCS are easy to find, but at the cost of the performance, compared to more general ATSs from Defn. 4.1 (e.g., RB #4 in Table 4.1).

4.5.2 The Unified NUM Problem

The NUM problem subject to UCS is given as follows:

$$\max_{\lambda_{\mathcal{A}}, x_{k\mathcal{C}}^{(\mathcal{A})}} \sum_{k \in \mathcal{U}} U\left(\sum_{\mathcal{A}} \sum_{\mathcal{C}:\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} r_{k\mathcal{C}}^{(\mathcal{A})}\right) \tag{4.5a}$$

$$\text{s.t.} \sum_{\mathcal{C}:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \frac{\sum_{k\in\mathcal{U}} x_{k\mathcal{C}}^{(\mathcal{A})}}{S_j(L_{\mathcal{A}})} \leq \lambda_{\mathcal{A}}, \ \forall j, \mathcal{A}, \tag{4.5b}$$

[4]The choice of $L_{\max}$ is a design choice. It depends on not only the average number of nearby BS arrays that users typically see but also the complexity that can be afforded.

$$\sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} \leq \lambda_{\mathcal{A}},\ \forall k, \forall \mathcal{A}, \tag{4.5c}$$

$$x_{k\mathcal{C}}^{(\mathcal{A})} \geq 0,\ \forall k, \forall \mathcal{C}, \forall \mathcal{A}, \tag{4.5d}$$

$$\sum_{\mathcal{A}} \lambda_{\mathcal{A}} \leq 1, \tag{4.5e}$$

$$\lambda_{\mathcal{A}} \geq 0,\ \forall \mathcal{A}, \tag{4.5f}$$

where the utility function $U(\cdot)$ is a continuously differentiable, monotonically increasing, and strictly concave function [49]. Constraint (4.5b) signifies that the total activity fractions allocated by BS j in clusters of size L cannot exceed the total available resources $\lambda_{\mathcal{A}} S_j(L_{\mathcal{A}})$. On the other hand, recalling that each user cannot be served by multiple clusters on the same RBs, (4.5c) signifies that the fraction of RBs over which user k is served by clusters in $\mathcal{A}$ cannot exceed RBs allocated for this ATM.

Note that (4.5) can also be applied to the schemes with each ATM consisted of disjoint clusters (i.e., the clusters without common BSs), by adapting $S_j(L_{\mathcal{A}})$ in (4.5b) to $S_j(|\mathcal{C}|)$. In this case, each BS belongs to at most one cluster on any RB. Considering the BS cluster as a new "super" BS, each ATM becomes exactly the same as the cellular transmission (i.e., no JT), and thus we can always find a feasible schedule to get the optimal performance as shown in [83]. However, the complexity to find all possible transmission modes with disjoint clusters is very high, which is the Bell number A_J with the recursion $A_{n+1} = \sum_{k=0}^{n} \binom{n}{k} A_k$ and $A_0 = 1$ in a network with J BSs [117]. Other general ATSs (e.g., asynchronous configuration in blanking approach and ATMs including clusters of different sizes in JT) may further improve the performance,

at the cost of complexity or intractability, and are left for future work.

Remark 4.1. *The problem formulation (4.5) is quite flexible, which can be applied to the joint optimization of user association and JT, as well as the joint optimization of user association and RB blanking. Note that JT and RB blanking are not exclusive. For example, we can adopt JT in both normal and blank RBs, and (4.5) is still applicable. Moreover, (4.5) can be applied regardless of how the resources are allocated among macro and small BSs (e.g., macro and small cells can either share or orthogonally use the resources).*

Remark 4.2. *The BS clusters in (4.5) are user specific. In other words, we consider the clustering of BSs from the perspective of each user, and thus different users can be served by different clusters.*

In this chapter, we specific the utility in (4.5) to the logarithmic utility as [30, 71, 83]. Logarithmic function as a very common choice of utility function, is well known as a objective function striking good balance between network throughput and user fairness [50]. It is easy to verify that (4.5) is a convex optimization problem [55]. General numerical solvers (e.g., CVX) can be used to solve (4.5). Since CVX is not well-suited for large instances [118], we alternatively propose an efficient algorithm that can be implemented in a partially distributed manner with low overhead in the next section.

4.6 Dual Subgradient Based Algorithm

In this section, we propose an efficient algorithm based on the dual-subgradient method. For the convergence analysis, we add a redundant variable $R_k = \sum_{\mathcal{A}} \sum_{\mathcal{C}:\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} r_{k\mathcal{C}}^{(\mathcal{A})}$ that changes the objective function in (4.5) to $\sum_k U(R_k)$, and redundant constraints $\sum_{\mathcal{A}} \sum_{\mathcal{C}} x_{k\mathcal{C}}^{(\mathcal{A})} \leq 1$ and $R_k \leq R_{\max}$ to (4.5), where $R_{\max} = \max_{k,\mathcal{A},\mathcal{C}} r_{k\mathcal{C}}^{(\mathcal{A})}$. We let θ_k, $\nu_{j\mathcal{A}}$ and $\mu_{k\mathcal{A}}$ be the Lagrange multipliers corresponding to $R_k = \sum_{\mathcal{A}} \sum_{\mathcal{C}:\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} r_{k\mathcal{C}}^{(\mathcal{A})}$, (4.5b) and (4.5c), respectively. Recalling that the number of ATMs in UCSs depends on the considered cluster sizes, the number of ATMs is $L_{\max}$. Denoting $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_K]^T$, $\boldsymbol{\nu} = \left[\boldsymbol{\nu}_{\mathcal{A}_1}^T, \cdots, \boldsymbol{\nu}_{\mathcal{A}_{L_{\max}}}^T\right]^T$ with $\boldsymbol{\nu}_{\mathcal{A}_i} = [\nu_{1\mathcal{A}_i}, \cdots, \nu_{N_B\mathcal{A}_i}]^T$, and $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_{\mathcal{A}_1}^T, \cdots, \boldsymbol{\mu}_{\mathcal{A}_{L_{\max}}}^T\right]^T$ with $\boldsymbol{\mu}_{A_i} = [\mu_{1\mathcal{A}_i}, \cdots, \mu_{K\mathcal{A}_i}]^T$, the dual problem of (4.5) is

$$\min_{\boldsymbol{\theta},\boldsymbol{\nu},\boldsymbol{\mu}\geq 0} \sum_{k\in\mathcal{U}} f_k(\boldsymbol{\theta},\boldsymbol{\nu},\boldsymbol{\mu}) + g(\boldsymbol{\nu},\boldsymbol{\mu}), \tag{4.6}$$

where

$$\begin{aligned} f_k(\boldsymbol{\theta},\boldsymbol{\nu},\boldsymbol{\mu}) = \max_{\substack{R_k\in[0,R_{\max}],\\ \sum_{\mathcal{A}}\sum_{\mathcal{C}} x_{k\mathcal{C}}^{(\mathcal{A})}\leq 1, x_{k\mathcal{C}}\geq 0}} & \log(R_k) - \theta_k R_k + \theta_k \sum_{\mathcal{A}} \sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} r_{k\mathcal{C}}^{(\mathcal{A})} \\ & + \sum_{\mathcal{A}} \sum_{j\in\mathcal{B}} \nu_{j\mathcal{A}} \sum_{\mathcal{C}:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \frac{x_{k\mathcal{C}}^{(\mathcal{A})}}{S_j(|\mathcal{C}|)} - \sum_{\mathcal{A}} \mu_{k\mathcal{A}} \sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}, \end{aligned} \tag{4.7}$$

and

$$g(\boldsymbol{\nu},\boldsymbol{\mu}) = \max_{\substack{\sum_{\mathcal{A}}\lambda_{\mathcal{A}}\leq 1,\\ \lambda_{\mathcal{A}}\geq 0}} \sum_{\mathcal{A}} \sum_{j:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \nu_{j\mathcal{A}}\lambda_{\mathcal{A}} + \sum_{\mathcal{A}} \sum_{k\in\mathcal{U}} \mu_{k\mathcal{A}}\lambda_{\mathcal{A}}. \tag{4.8}$$

The constraints of (4.5) satisfy the Slater condition [55], and thus the strong duality holds (i.e., the dual problem (4.6) and the original problem (4.5) have the same optimal value).

4.6.1 The Dual Subgradient Method

The optimization problem (4.7) can be viewed from the user's perspective, whose optimal solutions have closed-form: $R_k^* = \min\{1/\theta_k, R_{\max}\}$ and[5]

$$x_{k\mathcal{C}}^{(\mathcal{A})} = \begin{cases} 1, & \text{if } \{\mathcal{C},\mathcal{A}\} = \arg\max_{\mathcal{C},\mathcal{A}} \left(\theta_k r_{k\mathcal{C}}^{(\mathcal{A})} + \sum_{j:j\in\mathcal{C}} \frac{\nu_{j\mathcal{A}}}{S_j(|\mathcal{C}|)} - \mu_{k\mathcal{A}}\right), \\ 0, & \text{otherwise.} \end{cases} \tag{4.9}$$

The problem (4.8) is a linear programming and one optimal solution is

$$\lambda_{\mathcal{A}}^* = \begin{cases} 1, \text{ if } \mathcal{A} = \mathcal{A}^* \\ \\ 0, \text{ otherwise,} \end{cases} \tag{4.10}$$

where $\mathcal{A}^* = \arg\max_{\mathcal{A}} \sum_{j:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \nu_{j\mathcal{A}} + \sum_{k\in\mathcal{U}} \mu_{k\mathcal{A}}$[6].

The tth iteration of the dual subgradient algorithm is given as follows.

1. Associate user k to cluster $\mathcal{C}$ with $x_{k\mathcal{C}}^{(\mathcal{A})}(t) = 1$, where $x_{k\mathcal{C}}^{(\mathcal{A})}(t)$ are obtained by (4.9) with $\theta_k = \theta_k(t)$, $\nu_{j\mathcal{A}} = \nu_{j\mathcal{A}}(t)$ and $\mu_{k\mathcal{A}} = \mu_{k\mathcal{A}}(t)$.

2. Update the fraction of resources allocated to different ATMs by (4.10), where $\nu_{j\mathcal{A}} = \nu_{j\mathcal{A}}(t)$ and $\mu_{k\mathcal{A}} = \mu_{k\mathcal{A}}(t)$.

3. Update the Lagrangian multipliers by

$$\theta_k(t+1) = \theta_k(t) - \delta(t) \left(\sum_{\mathcal{A}} \sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}(t) r_{k\mathcal{C}}^{(\mathcal{A})} - \min\{1/\theta_k(t), R_{\max}\} \right), \tag{4.11}$$

[5]If we have multiple pairs of $\{\mathcal{C},\mathcal{A}\}$ that maximize $\left(\theta_k r_{k\mathcal{C}}^{(\mathcal{A})} + \sum_{j:j\in\mathcal{C}} \frac{\nu_{j\mathcal{A}}}{S_j(|\mathcal{C}|)} - \mu_{k\mathcal{A}}\right)$, we just randomly pick one pair.

[6]If we have multiple $\mathcal{A}$ that maximize the $\sum_{j:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \nu_{j\mathcal{A}} + \sum_{k\in\mathcal{U}} \mu_{k\mathcal{A}}$, we just randomly pick one of them and let it be $\mathcal{A}^*$.

$$\nu_{j\mathcal{A}}(t+1) = \left[\nu_{j\mathcal{A}}(t) - \delta(t)\left(\lambda_{\mathcal{A}}(t) - \sum_{\mathcal{C}:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \frac{\sum_{k\in\mathcal{U}} x_{k\mathcal{C}}^{(\mathcal{A})}(t)}{S_j(|\mathcal{C}|)}\right)\right]^+, \quad (4.12)$$

and $\mu_{k\mathcal{A}}(t+1) = \left[\mu_{k\mathcal{A}}(t) - \delta(t)\left(\lambda_{\mathcal{A}}(t) - \sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}(t)\right)\right]^+$, where $[z]^+ = \max\{z, 0\}$ and $\delta(t)$ is the stepsize at tth iteration.

Proposition 4.3. *The dual subgradient algorithm converges to the optimal Lagrangian multipliers, if the stepsize $\delta(t)$ is chosen appropriately (e.g., the diminishing stepsize $\delta(t) = a/(t+b)$, where a and b are some positive scalars).*

Proof. Based on Prop. 6.3.4. in [55], we complete the proof by showing that subgradients $\lambda_{\mathcal{A}} - \sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}$, $\sum_{\mathcal{A}}\sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}(t) r_{k\mathcal{C}}^{(\mathcal{A})} - \min\{\frac{1}{\theta_k(t)}, R_{\max}\}$ and $\lambda_{\mathcal{A}} - \sum_{\mathcal{C}:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \frac{\sum_{k\in\mathcal{U}} x_{k\mathcal{C}}^{(\mathcal{A})}}{S_j(|\mathcal{C}|)}$ are bounded. □

4.6.2 Finding the Optimal Primal Solutions Given Optimal Lagrangian Multipliers

Note that the objective function of (4.5) is not strictly convex and we may have multiple optimal solutions. In general, given the optimal dual variables, it is difficult to find the optimal primal solutions that satisfy the KKT conditions. However, by exploring the structure of (4.5) as follows, we propose to obtain the optimal primal solutions by solving a small-size LP.

The optimal long-term rate R_k^* in (4.5) is unique, since the function $\log(R_k)$ is strictly concave with respect to R_k. Thus, given the optimal dual variables, we can easily get the unique optimal rate

$$R_k^* = \max_{\mathcal{C},\mathcal{A}} \left\{ \frac{r_{k\mathcal{C}}^{(\mathcal{A})}}{\sum_{j\in\mathcal{C}} \nu_{j\mathcal{A}}/S_j(L_{\mathcal{A}}) + \mu_{k\mathcal{A}}} \right\}$$

by exploring the KKT conditions of problem (4.5), which imply

$$R_k \geq \frac{r_{k\mathcal{C}}^{(\mathcal{A})}}{\sum_{j\in\mathcal{C}} \nu_{j\mathcal{A}}/S_j(L_{\mathcal{A}}) + \mu_{k\mathcal{A}}}. \tag{4.13}$$

We can observe from (4.13) that in the optimal solutions, each user only has positive activity fractions to clusters providing the maximum $\frac{r_{k\mathcal{C}}^{(\mathcal{A})}}{\sum_{j\in\mathcal{C}} \nu_{j\mathcal{A}}/S_j(L_{\mathcal{A}})+\mu_{k\mathcal{A}}}$. Leveraging this conclusion, we propose the following LP, whose size is reduced by only focusing on the positive $x_{k\mathcal{C}}^{(\mathcal{A})}$ obtained from (4.13).

$$\begin{aligned}
\max_{\eta,x,\lambda}\ & \eta \\
\text{s.t. } & \eta \leq \sum_{\mathcal{A}} \sum_{\mathcal{C}\in\mathcal{A}} \frac{x_{k\mathcal{C}}^{(\mathcal{A})} r_{k\mathcal{C}}^{(\mathcal{A})}}{R_k^*},\ \forall k \in \mathcal{U}, \\
& \sum_{\mathcal{C}: j\in\mathcal{C}, \mathcal{C}\in\mathcal{A}} \frac{\sum_{k\in\mathcal{U}} x_{k\mathcal{C}}^{(\mathcal{A})}}{S_j(L_{\mathcal{A}})} \leq \lambda_{\mathcal{A}},\ \forall j, \forall \mathcal{A}, \\
& \sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} \leq \lambda_{\mathcal{A}},\ \forall k, \forall \mathcal{A}, \\
& \sum_{\mathcal{A}} \lambda_{\mathcal{A}} \leq 1, \\
& x_{k\mathcal{C}}^{(\mathcal{A})}, \lambda_{\mathcal{A}} \geq 0,\ \forall k, \forall \mathcal{C}, \forall \mathcal{A}.
\end{aligned} \tag{4.14}$$

Proposition 4.4. *Given that R_k^* is the exact optimal rate of (4.5), the solutions of (4.14) are the same as the optimal solutions of problem (4.5).*

Proof. Similar techniques in the proof of Lemma 1 in [83] can be used to complete this proof. □

Prop. 4.4 implies that we can obtain the optimal primal variables as long as $\nu_{j\mathcal{A}}^*$ and $\mu_{k\mathcal{A}}^*$ from the dual subgradient algorithm are optimal. Though Prop. 4.3 shows the convergence of the dual subgradient algorithm, there

may exist a small gap between the obtained dual variables and the optimal solutions, due to the numerical precision or the limited number of iterations. Exploiting the well-behaved structure of LP (4.14), i.e., finite coefficients and a bounded feasible set [83], it is expected that the solution of (4.14) is still near optimal in the presence of a small gap from optimal dual variables.

4.6.3 Implementation Discussion

The above algorithm can be implemented in either a centralized or a partially distributed manner, where the first part of the algorithm (i.e., steps (4.9)-(4.12)) to get optimal dual variables can be implemented distributively. To discuss the complexity and overhead, we first explore the properties of optimal solutions. Empirical evidence reveals that in a heavily loaded network, where the constraints (4.5c) are inactive (i.e., $\sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} < \lambda_{\mathcal{A}}$), most users are uniquely served by a cluster per ATM. Insight regarding this observation can be obtained by examining KKT conditions of (4.5).

Proposition 4.5. *For a given $\mathcal{A}$, if (4.5c) are inactive $\forall k \in \mathcal{U}$, the number of users that are served by multiple BS clusters in $\mathcal{A}$ is at most $N_{\mathcal{C}\mathcal{A}} - 1$, where $N_{\mathcal{C}\mathcal{A}}$ is the number of clusters in $\mathcal{A}$.*

Proof. See Appendix 4.11.4. □

Prop. 4.5 implies that the user association in NUM solutions are mostly unique (i.e., users are served by a single cluster per ATM). Note that $N_{\mathcal{C}\mathcal{A}}$ provides an upper bound, and the number of fractional users is much smaller

than $N_{\mathcal{C}\mathcal{A}}$ in simulations (e.g., less than 3.5% in Sec. 4.9). Thus, given the optimal dual variables, there are limited number of users have multiple optimal solutions to (4.9). The activity fractions of users with unique association can be solved by (4.13) with equality directly. The unknown activity fractions that needs to be solve via (4.14) are only the ones of fractional users in the active ATMs (i.e., the $\mathcal{A}$ with $\lambda_{\mathcal{A}} > 0$, that maximize $\sum_{j:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \nu_{j\mathcal{A}} + \sum_{k\in\mathcal{U}} \mu_{k\mathcal{A}}$). As shown below, this observation enables the efficiency and the significantly reduced overhead of the above proposed algorithm.

Complexity discussion. Let L_a and $N_{\mathcal{C}m}$ be the number of active ATMs and the number of clusters in the active ATM that has the most clusters (i.e., $N_{\mathcal{C}m} = \max_{\mathcal{A}:\lambda_{\mathcal{A}}>0} N_{\mathcal{C}\mathcal{A}}$), respectively. The size of LP (4.14) is $O(N_{\mathcal{C}m} L_a \min\{N_{\mathcal{C}m} - 1, K\})$, which is smaller than the size of the original nonlinear NUM problem (4.5) that is $O(N_{\mathcal{C}m} L_{\max} K)$. As shown in Sec. 4.9, the number of fractional users is very small (less than $3.5\%K$), and thus the size of (4.14) is much smaller (less than 3.5%) than (4.5). The size of LP can be further reduced when $L_a/L_{\max}$ is small (e.g., only 2 ATMs are active among 4 ATMs in Fig. 4.9). Note that we can either use (4.14) or (4.5) with the reduced size to get these positive activity fractions. Thus, the complexity of the second part (i.e., the LP) of the algorithm can be correspondingly reduced by reducing the problem size (to less than 3.5% of the original size), at the cost of additional computation in steps (4.9)-(4.12), whose complexity is $O(L_{\max} N_{\mathcal{C}m})$ per user at each iteration, which is much smaller than $O(L_{\max} N_{\mathcal{C}m} K)$. Recalling that each user can conduct computation in parallel if the algorithm runs

in a partially distributed manner, the low complexity per user in the first part (i.e., steps (4.9)-(4.12)) of the algorithm and the reduced problem size in the second part of the algorithm imply the efficiency of the proposed algorithm.

Overhead discussion. When the algorithm is implemented in a partially distributed manner, denoting by N the number of required iterations for the dual subgradient method that is quite small as shown in the simulation (about 60), the overhead required by the proposed algorithm is $O\left((K+JL_{\max})N+L_aN_{\mathcal{C}m}\min\{N_{\mathcal{C}m}-1,K\}\right)$, where the first term $(K+JL_{\max})N$ corresponds to the first part of the proposed algorithm (i.e., dual subgradient method) and the second term $L_aN_{\mathcal{C}m}\min\{N_{\mathcal{C}m}-1,K\}$ corresponds to the overhead required for LP (4.14). Compared to the overhead in centralized manners that is $O(KN_{\mathcal{C}m}L_{\max})$, the overhead is significantly reduced. Taking our simulation for example, there are 840 users, 255 potential clusters serving each user, and 34 users with multiple $\{\mathcal{C}^*,\mathcal{A}^*\}$ where the number of multiple $\{\mathcal{C}^*,\mathcal{A}^*\}$ is less than 50 per user. Thus, the overhead is reduced from about 4.3×10^5 to 6.2×10^4 (reduced more than 85%).

As a result, we propose an efficient low-overhead algorithm that can be implemented in a partially distributed manner. Note that though the optimal activity fractions $x_{k\mathcal{C}}^{*(\mathcal{A})}$ are obtained, it is not *a prior* known whether the optimal solutions can be implemented via any schedulers or not. The implementation of the optimal solutions is discussed in the following section.

4.7 Virtual Queue Based Scheduling Scheme

In this section we aim to propose scheduling policies that yield activity fractions $\left\{x_{k\mathcal{C}}^{(\mathcal{A})}\right\}$ closely matching the solution of (4.5). We first introduce the definition of scheduling policies as follows.

Definition 4.3. Feasible Schedule: *A scheduling policy* $\left\{\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t);\ \forall \mathcal{A}, \forall \mathcal{C},\ \forall t \in \{1, 2 \cdots, T\}\right\}$ *is feasible with respect to the ATSs based on Defn. 4.2 if it satisfies the following:*

(i) For each t, the policy associates with RB t one ATM $\mathcal{A}$ with clusters of size $L_{\mathcal{A}}$; that is, for each cluster $\mathcal{C}$ with $\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t)$ being non-empty, we have $|\mathcal{C}| = L_{\mathcal{A}}(t)$.

(ii) For each t, each user is served by at most one cluster; that is, $|\sum_{\mathcal{C}} \mathbb{1}\{k \in \mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t)\}| \leq 1$.

(iii) For each t allocated to $\mathcal{A}$ and for each BS j, BS j serves at most $S_j(L_{\mathcal{A}}(t))$ users; that is, $|\cup_{\mathcal{C}:\, j\in\mathcal{C}} \mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}(t)| \leq S_j(L_{\mathcal{A}}(t))$.

4.7.1 The Feasibility of the NUM Solution in Implementation

It is easy to verify that $\left\{x_{k\mathcal{C}}^{(\mathcal{A})}\right\}$ and $\{\lambda_{\mathcal{A}}\}$ yielded by any feasible schedules defined by Defn. 4.3 satisfy (4.5b)-(4.5d). However, there exist $\left\{x_{k\mathcal{C}}^{(\mathcal{A})}\right\}$ and $\{\lambda_{\mathcal{A}}\}$ satisfying (4.5b)-(4.5d), for which no feasible schedule in the sense of Defn. 4.3 exists.

Theorem 4.1. *In UCSs with $L_{\max} > 1$, there exist some activity fractions satisfying (4.5b)-(4.5d) that cannot be implemented by any feasible schedule in*

Defn. 4.3. If $L_{\max} = 1$, then there exists at least one feasible schedule that can provide long-term activity fractions approaching the optimal solution of (4.5).

Proof. See Appendix 4.11.3. □

Finding the constraints to make any activity fractions satisfying these constraints be realizable by a feasible schedule is essentially finding an integral hull explicitly in the convex hull characterized by (4.5b)-(4.5d), which is very difficult if not impossible. Though some solutions of (4.5) cannot be realized by feasible schedules, they provide upper bounds on the network performance, which can serve as benchmarks. The infeasibility of NUM solutions motivates us to develop a scheduling scheme that can provide approximate but implementable activity fractions.

4.7.2 The Greedy Virtual Queue Scheduling Scheme

We consider scheduling policies for the UCS comprised of $L_{\max}$ parallel schedulers, one per ATM $\mathcal{A}$. We describe a method for scheduling users over the RBs in the $\lambda_{\mathcal{A}} > 0$ fraction that are dedicated to clusters $\mathcal{C} \in \mathcal{A}$.

Observing that most users have unique association (implied by Prop. 4.5), we approximate the NUM solution to the unique association per ATM (e.g., let the cluster providing the largest activity fraction be the only cluster serving the user on each RB), where the approximate activity fraction, denoted

by $\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})}$, is given by

$$\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})} = \begin{cases} x_{k\mathcal{C}}^{(\mathcal{A})} & \text{if } \mathcal{C} = \arg\max_{\mathcal{C}:\ \mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} \\ 0 & \text{otherwise} \end{cases}. \tag{4.15}$$

Let $\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}$ denote the set of users for which $\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})} > 0$. In the unique association, we have $\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})} \cap \mathcal{U}_{\mathcal{C}'}^{(\mathcal{A})} = \emptyset, \forall \mathcal{C} \neq \mathcal{C}', \mathcal{C} \in \mathcal{A}, \mathcal{C}' \in \mathcal{A}$. To assign user k a fraction of RBs close to the desired fraction $\alpha_k^{*(\mathcal{A})} = \frac{\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})}}{\lambda_{\mathcal{A}}}$, we consider the max-min scheduling, whose objective function is $\max\min_k \sum_t \alpha_k^{(\mathcal{A})}(t)\tilde{R}_k^{(\mathcal{A})}$, where $\tilde{R}_k^{(\mathcal{A})} = 1/\alpha_k^{*(\mathcal{A})}$. We ignore the cluster index $\mathcal{C}$ in $\tilde{R}_k^{(\mathcal{A})}$ given the unique association. The metric $\tilde{R}_k^{(\mathcal{A})}$ can be considered as the hypothetical peak-rate of user k from cluster $\mathcal{C}$ in $\mathcal{A}$.

If $\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})}$ can be implemented by a feasible schedule, any scheduling schemes for max-min fairness can be adopted to get the average fraction of resources $\bar{\alpha}_k^{(\mathcal{A})} = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} \alpha_k^{(\mathcal{A})}(t)}{T} = \frac{1}{\tilde{R}_k^{(\mathcal{A})}} = \alpha_k^{*(\mathcal{A})}$. Thus, we change the implementation of activity fractions to the max-min fairness scheduling problem. In this chapter, we propose to solve the scheduling problem based on the virtual queue (VQ) method [119], where the scheduler performs the following weighted sum rate maximization at t [120]:

$$\max_{\left\{\alpha_k^{(\mathcal{A})}(t)\right\}} \sum_{\mathcal{C}\in\mathcal{A}} \sum_{k\in\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}} \alpha_k^{(\mathcal{A})}(t) Q_k^{(\mathcal{A})}(t) \tilde{R}_k^{(\mathcal{A})}, \tag{4.16a}$$

$$\text{s.t.} \sum_{\mathcal{C}:j\in\mathcal{C}} \sum_{k\in\mathcal{U}_{\mathcal{C}}^{(\mathcal{A})}} \frac{\alpha_k^{(\mathcal{A})}(t)}{S_j(L_{\mathcal{A}})} \leq 1, \tag{4.16b}$$

$$\alpha_k^{(\mathcal{A})}(t) \in \{0, 1\}. \tag{4.16c}$$

Parameter $Q_k^{(\mathcal{A})}(t)$ is the length of the VQ at user k in $\mathcal{A}$ at time t, which is updated by

$$Q_k^{(\mathcal{A})}(t+1) = \max\{0, Q_k^{(\mathcal{A})}(t) - \tilde{R}_k^{(\mathcal{A})}(t)\} + \Lambda_k^{(\mathcal{A})}(t), \tag{4.17}$$

where

$$\tilde{R}_k^{(\mathcal{A})}(t) = \begin{cases} \tilde{R}_k^{(\mathcal{A})}, \text{ if user } k \text{ is scheduled at time } t, \\ 0, \text{ otherwise,} \end{cases}, \tag{4.18}$$

with $\Lambda_{\max}$ and V chosen sufficiently large [119, 120]. Similar to Theorems 1 and 2 in [119], if $\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})}$ is realizable by some schedulers, we can show that the obtained time averaged activity fractions $\bar{\alpha}_k^{(\mathcal{A})}$ can be arbitrarily close to the optimal solution $\alpha_k^{*(\mathcal{A})}$ by letting V be as large as desired. On the other hand, for $\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})}$ that cannot be realized via any feasible schedule, though there is no guarantee on the maximum gap between $\bar{\alpha}_k^{(\mathcal{A})}$ and $\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})}$, $\bar{\alpha}^{(\mathcal{A})}$ can be very close to the feasible activity fractions whose element-wise gap from $\tilde{x}_{k\mathcal{C}}^{(\mathcal{A})}$ is minimized.

Scheduling via (4.16) is impractical, as it needs to solve an integer optimization (4.16) at each t. Alternatively, a number of heuristic algorithms can be used to provide feasible (though generally suboptimal) solutions. In this chapter, we consider a rudimentary greedy algorithm. Letting $K_{\mathcal{A}} = |\cup_{\mathcal{C}\in\mathcal{A}}\mathcal{U}_{\mathcal{C}}|$, the greedy algorithm operates at time t as follows:

1. Determine a user order $\pi(k)$, for which $Q_{\pi(k)}^{(\mathcal{A})}(t)\tilde{R}_{\pi(k)}^{(\mathcal{A})} \geq Q_{\pi(k+1)}^{(\mathcal{A})}(t)\tilde{R}_{\pi(k+1)}^{(\mathcal{A})}$ for all k.

2. Initialization: $k = 1$ and $\tilde{\mathcal{U}} = \emptyset$.

3. If the user set $\tilde{\mathcal{U}} \cup \{\pi(k)\}$ satisfies all the constraints in (4.16), set $\tilde{\mathcal{U}} = \tilde{\mathcal{U}} \cup \{\pi(k)\}$ and correspondingly update $\alpha^{(\mathcal{A})}_{\pi(k)\mathcal{C}}(t)$.

4. If $k < K_{\mathcal{A}}$ and there exists at lease one BS with (4.16b) being a strict inequality, set $k = k + 1$ and go to step 3.

5. Output $\tilde{\mathcal{U}}$ as the scheduling user set at time t.

4.8 Possible Alterantive Method – the Slot-based Gradient Algorithm [1]

As introduced in Section 4.1, an alternative method to solve the joint clustering and resource allocation problem is to leverage the slot-based gradient algorithm proposed in [1]. Recalling that the NUM solutions of the UCS scheme in Defn. 4.2 may not be realizable by any feasible schedulers, the proposed greedy VQ scheduling scheme can be considered as an algorithm to approximate the outer-bound solution (i.e., the NUM solution that may be outside of the realizable region) to a sub-optimal inner-bound solution (i.e., the solutions in the realizable region). In contrast, the slot-based framework tackle this problem from the inner bound, as the solution at each slot is realizable. In this section, we give a first-cut investigation in applying the slot-based framework to the considered problem.

Let $\bar{R}_k$ be the long-term rate of user k. At time slot t, the system chooses the scheduling decision $x^{(\mathcal{A})}_{k\mathcal{C}}(t)$ and $\lambda_{\mathcal{A}}(t)$ by solving the following max-

imization problem

$$\max_{\lambda_{\mathcal{A}}(t),x_{k\mathcal{C}}^{(\mathcal{A})}(t)} \sum_{k\in\mathcal{U}} \frac{\partial U(\bar{R}_k)}{\partial \bar{R}_k}\bigg|_{\bar{R}_k(t)} R_k(t) \tag{4.19a}$$

$$\text{s.t.} \sum_{\mathcal{C}:j\in\mathcal{C},\mathcal{C}\in\mathcal{A}} \frac{\sum_{k\in\mathcal{U}} x_{k\mathcal{C}}^{(\mathcal{A})}(t)}{S_j(|\mathcal{C}|)} \leq \lambda_{\mathcal{A}}(t),\ \forall j,\mathcal{A}, \tag{4.19b}$$

$$\sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}(t) \leq \lambda_{\mathcal{A}}(t),\ \forall k,\mathcal{A}, \tag{4.19c}$$

$$\sum_{\mathcal{A}} \lambda_{\mathcal{A}}(t) \leq 1, \tag{4.19d}$$

$$x_{k\mathcal{C}}^{(\mathcal{A})}(t), \lambda_{\mathcal{A}}(t) \in \{0,1\},\ \forall k,\mathcal{C},\mathcal{A}, \tag{4.19e}$$

where $\bar{R}_k(t+1) = (1-\beta)\bar{R}_k(t) + \beta R_k(t)$, and $R_k(t) = \sum_{\mathcal{A}} \sum_{\mathcal{C}:\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})}(t) r_{k\mathcal{C}}^{(\mathcal{A})}$. According to the conclusion in [1], we have that as $\beta \to 0$, both the estimate rate $\lim_{t\to\infty} \bar{R}_k(t)$ and the long-term rate $\bar{R}_k$ obtained by solving (4.19) at each slot, converge to the optimal long-term rate $\bar{R}_k^*$.

The complexity to solve (4.19) is very high (e.g., exponential complexity using the brute force search). Low-complexity efficient (though generally sub-optimal) algorithms that can be applied to each slot are needed. We propose a greedy algorithm, which operates at time t as follows:

1. Initialization: $k = 1$, $\tilde{\mathcal{U}}^{(\mathcal{A})} = \emptyset$ and $l_{\mathcal{A}} = 1, \forall \mathcal{A}$;

2. For each $\mathcal{A}$, denote by $\pi_{\mathcal{A}}(l)$ the decreasing order in terms of $r_{k\mathcal{C}}^{(\mathcal{A})}/\bar{R}_k(t)$, for which

$$r_{k_{\pi_{\mathcal{A}}(l)}\mathcal{C}_{\pi_{\mathcal{A}}(l)}}^{(\mathcal{A})}/\bar{R}_{k_{\pi_{\mathcal{A}}(l)}}(t) \geq r_{k_{\pi_{\mathcal{A}}(l)}\mathcal{C}_{\pi_{\mathcal{A}}(l+1)}}^{(\mathcal{A})}/\bar{R}_{k_{\pi_{\mathcal{A}}(l+1)}}(t);$$

3. For each $\mathcal{A}$, if user $k_{\pi_{\mathcal{A}}(l_{\mathcal{A}})}$ has not been scheduled, and all BSs in $\mathcal{C}_{\pi_{\mathcal{A}}(l_{\mathcal{A}})}$ have available resources (i.e., the activity fractions of set $\tilde{\mathcal{U}}^{(\mathcal{A})} \cup \{k_{\pi_{\mathcal{A}}(l_{\mathcal{A}})}\}$ satisfies all constraints in (4.19)), let $\tilde{\mathcal{U}}^{(\mathcal{A})} = \tilde{\mathcal{U}}^{(\mathcal{A})} \cup \{k_{\pi_{\mathcal{A}}(l_{\mathcal{A}})}\}$, and update $x^{(\mathcal{A})}_{k_{\pi_{\mathcal{A}}(l_{\mathcal{A}})}\mathcal{C}_{\pi_{\mathcal{A}}(l_{\mathcal{A}})}}(t)$ correspondingly;

4. If there exists a ATM $\mathcal{A}$ that $\tilde{\mathcal{U}}^{(\mathcal{A})} \neq \mathcal{U}$ and $l_{\mathcal{A}} < KN_{\mathcal{C}\mathcal{A}}$, set $l_{\mathcal{A}} = l_{\mathcal{A}} + 1$ and go to step 3;

5. Choose the ATM $\mathcal{A}$ with the largest $\left(\sum_k \sum_{\mathcal{C}\in\mathcal{A}} x^{(\mathcal{A})}_{k\mathcal{C}}(t) r^{(\mathcal{A})}_{k\mathcal{C}}\right) / \bar{R}_k(t)$, denoted by $\mathcal{A}^*$, and output $\tilde{\mathcal{U}}^{(\mathcal{A}^*)}$ as the set of scheduled users at time t.

Though the complexity of the above slot-based greedy algorithm (which is $O(N_{\mathcal{A}} N_{\mathcal{C}m} K \log(K))$ at each slot) and the complexity of the greedy VQ scheme (which is $O(K \log(K))$ at each slot) are comparable, the utility obtained using the slot-based greedy algorithm is much less than the utility obtained in the outer-bound approach proposed in this chapter, as shown in Fig. 4.5 in Section 4.9. We leave the investigation of other algorithms in the slot-based framework for future work.

4.9 Performance Evaluation

In this section, we present a simulation-based evaluation based on the "wrap-around" layout in Fig. 4.2. The parameters used are given as follows unless otherwise specified. There are 4 macros with $M_j = 100$ and $S_j(|\mathcal{C}|) = 10\rho|\mathcal{C}|$, and 32 pico BSs with $M_j = 40$ and $S_j(|\mathcal{C}|) = 4\rho|\mathcal{C}|$, with ρ being a tunable parameter in $[0, 1]$. One pico BS is at the center of each white square,

while 3 pico BSs are dropped uniformly within each shaded square (hotspot). Also, 15 and 90 single-antenna users are dropped uniformly in each white and each shaded square, respectively. The macro and pico BS transmit powers are 46dBm and 35dBm, respectively. The path-loss for macro-user links and pico-user links are $128.1 + 37.6 \log_{10} d$ and $140.7 + 36.7 \log_{10} d$, respectively, with the distance d in km. The noise power spectral density is -174 dBm/Hz. The largest cluster size considered is $L_{\max} = 4$.

We consider two distinct macro-pico operation scenarios: (i) macros and picos operate on the same band; (ii) macros and picos operate on different bands, where we provide macros a fraction of 20% RBs as an illustrative example, and macros can only engage in cellular transmission. Note that though we can jointly optimize the resource partition and user association using (4.5), we consider predefined RB partition for (ii), which provides a lower bound and serves as a benchmark. This is due to that resource partition among macro and small BSs is most likely static (or semi-static) in practice. Also, the macro and small BSs may operate on different frequency bands (e.g., the macro BSs may transmit on lower-frequency bands, while the small BSs may transmit on higher-frequency bands), where the amount of spectrum depends on the available bands and thus is not a variable to be optimized.

There exists a one-one mapping between the log utility and the geometric mean of rates as $\left(\prod_{k=1}^{K} R_k\right)^{1/K} = \exp\left(\frac{1}{K}\sum_{k=1}^{K} \log R_k\right)$, thus we use the geometric mean of rates as the metric for performance evaluation. Figs. 4.3 and 4.4 show the geometric mean of rates using different approaches. The

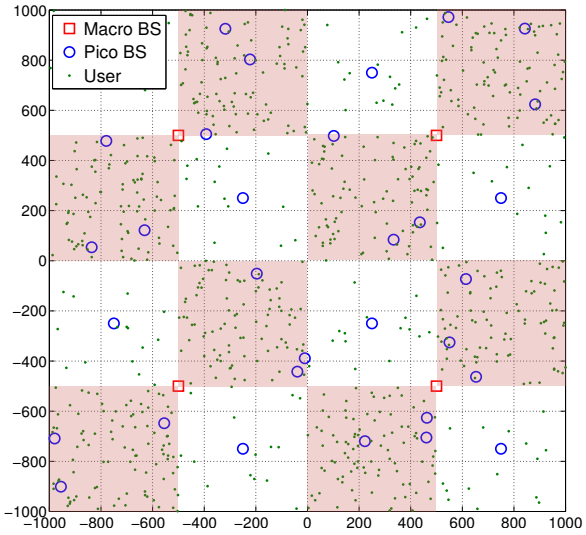


Figure 4.2: The illustration of network deployment. The white grids are the regular areas, while the shadowed grids are hotspots.

NUM solution refers to the solution of (4.5) solved by CVX. In Fig. 4.3, we can observe that JT significantly improves the geometric mean of rates versus the case with optimal user association but without interference management (i.e., the cellular transmission) and with the max-SINR association, in both macro-pico shared and orthogonal operation scenarios (about 1.6x in the shared case and 1.35x in the orthogonal case). The algorithm based on dual subgradient method has almost the same performance as the NUM solution, which validates our analysis. The algorithm based on dual subgradient method has almost the same performance as the NUM solution, which validates our analysis. The proposed greedy VQ scheduling scheme provides performance close the NUM solution, within 90% of the utility provided by NUM solutions in both the shared and orthogonal operation scenarios. Though in our setup, the

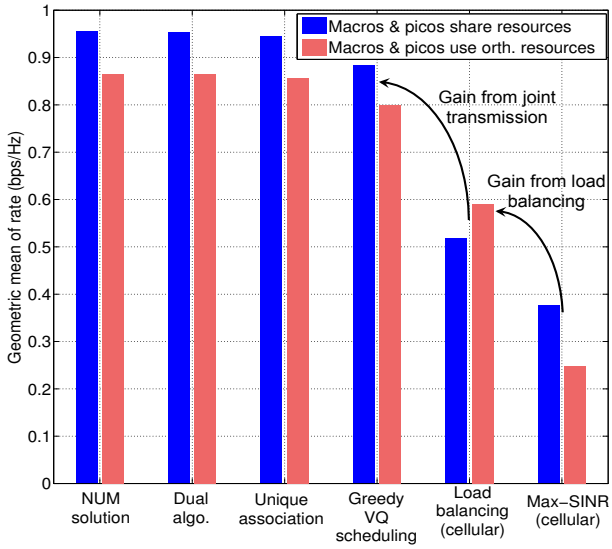


Figure 4.3: The geometric mean of rates using different approaches, when $\rho = 1$.

orthogonal operation scenario in cellular transmission outperforms the shared operation scenario, which one performs better depends heavily on the parameters such as transmit power and BS densities and it is not *a prior* known in general cases. Similar observations exist in Fig. 4.4, where the dual subgradient based algorithm and greedy VQ scheduling scheme approach the NUM solution with RB blanking. Note that we only adopt RB blanking in the cases where macro and pico BSs share the resources. We observe that RB blanking further improves the network utility. The gap between the NUM solution and the results of greedy VQ scheduling scheme is much smaller in cellular transmission than the cases with multiple-BS clusters (i.e., JT with $L_{\max} > 1$).

Fig. 4.5 compares the performance of the proposed greedy VQ scheme

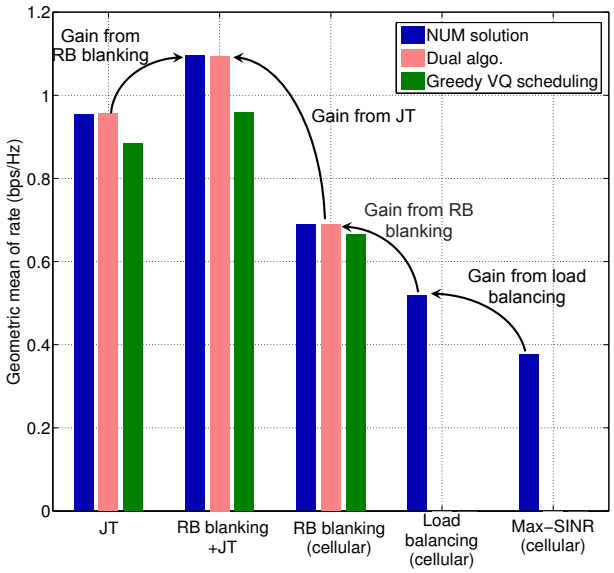


Figure 4.4: The geometric mean of rates with RB blanking, when $\rho = 1$ and macro and pico BSs share the resources. Blanking further improves the network performance.

to the result of the greedy algorithm in the slot-based framework. We observe that the greedy VQ scheme provides solutions that are quite close to the NUM solution, while the greedy slot-based algorithm has much less utility than the greedy VQ scheme. We leave the investigation of other algorithms that may have potential better performance in the slot-based framework for future work.

We observe similar conclusions in Fig. 4.6, which shows the rate cumulative distribution function (CDF) with different approaches. Note that the rate here refers to long-term rate, which incorporating the resource sharing among users served by the same BS. The rate of bottom (the 10th percentile) users with JT is about 2.2x of the case with optimal user association but without interference management (i.e., load balanced cellular transmission).

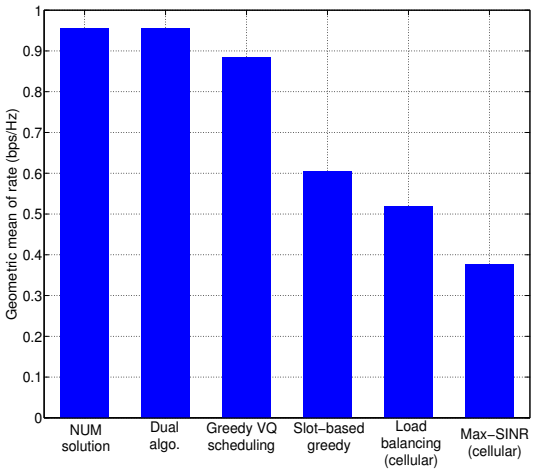
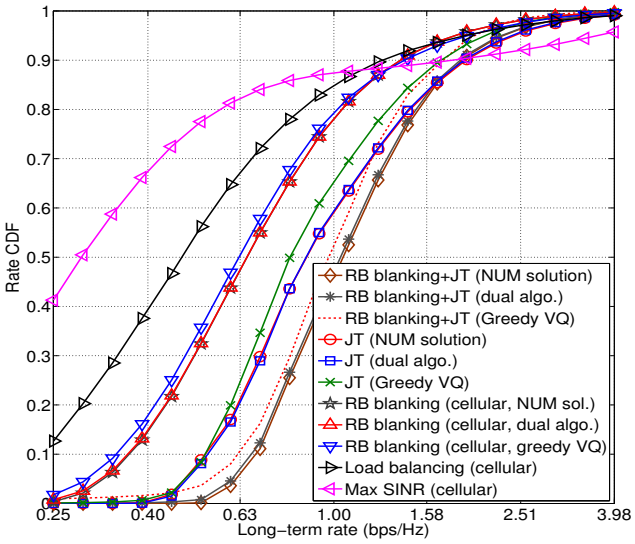


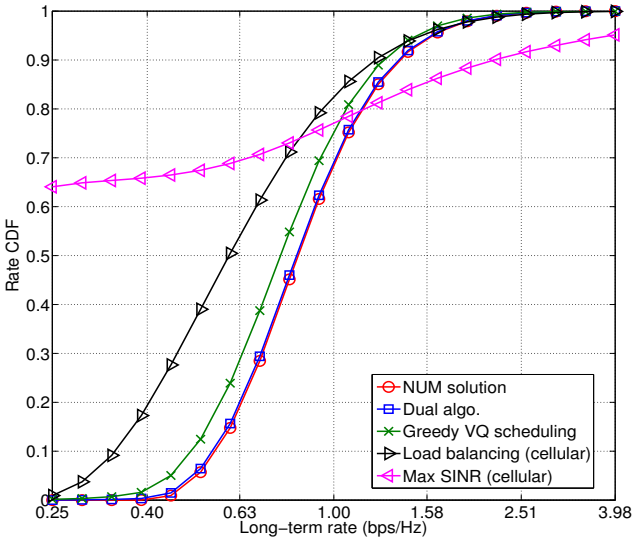
Figure 4.5: The comparison between the geometric mean of long-term rates using the slot-based framework with greedy algorithm and the greedy VQ scheme, when $\rho = 1$ in cases where macro and pico BSs share resources. The utility obtained from the greedy algorithm in the slot-based framework is much less the utility obtained from the greedy VQ scheme.

To further validate the performance of the greedy VQ scheduling scheme, we present the rate differences between the greedy VQ scheduling scheme and the NUM solution in Fig. 4.7. We can observe that there are 82.62% users whose rate differences between the greedy VQ scheduling scheme and the NUM solution are within 10% of the NUM solution. This implies that the greedy VQ scheduling scheme performs quite well.

The number of users served by different clusters with JT is illustrated in Fig. 4.8. In the max-SINR association, most users connect to the macro BSs, since macro BSs have larger transmit power. By load balancing, many users are offloaded to small BSs in the cellular transmission case with optimal user association. In our proposed framework, all users are served by BS



(a) Macro and pico BSs share resources



(b) Orthogonal resources for macro and pico BSs

Figure 4.6: The long-term rate CDF using different approaches, when $\rho = 1$.

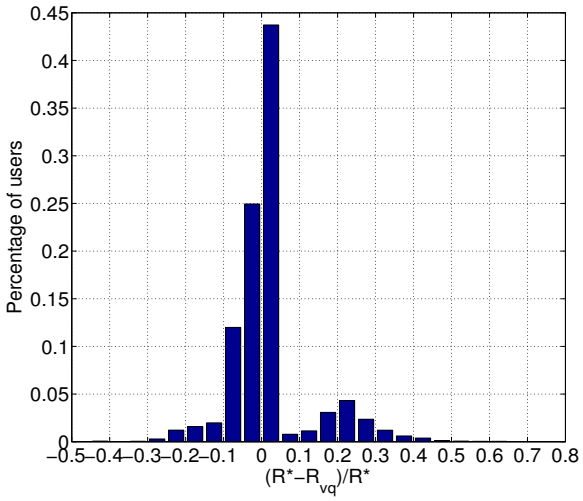
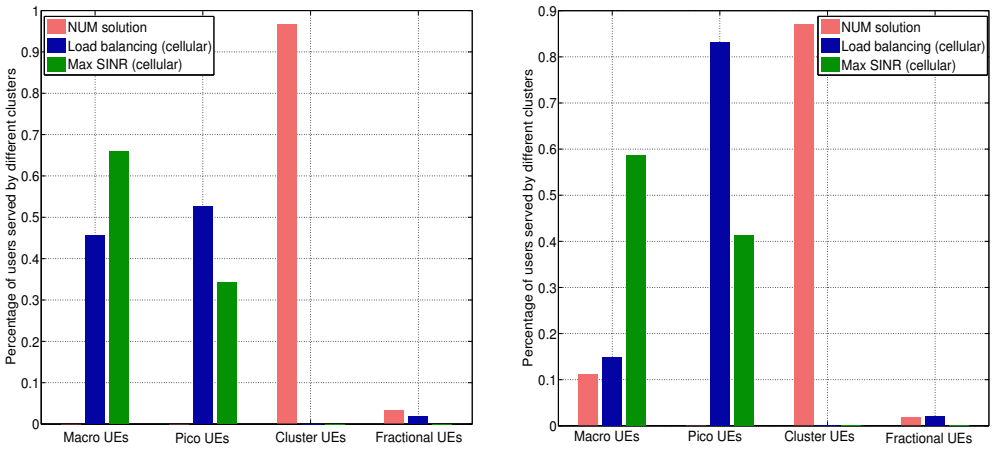


Figure 4.7: The rate differences between the greedy VQ scheduling scheme and the NUM solution, when $\rho = 1$ and macro and pico BSs share resources. There are 82.62% users whose rate differences between the greedy VQ scheduling scheme and the NUM solution are within 10% of the NUM solution.

clusters with multiple BSs, which implies the potential gain using JT. In the orthogonal resource allocation, more users connect to small BSs in the max-SINR association, since there are no cross-tier interference and more users may get larger SINR from small BSs than macro BSs. Due to the limited resources (20% RBs) available in macro BSs, more users are offloaded to small BSs using the load balancing approach in orthogonal cellular transmission case. The percentage of fractional users (i.e., users served by multiple clusters) is about 3.3% using JT, and 1.2% in the case without JT. For blanking, the percentage of fractional users in the case using JT with RB blanking is about 2.5%, while the percentage of fractional users in the case with blanking but without JT is less than 1%. Thus, we can conclude that the number of fractional users in all



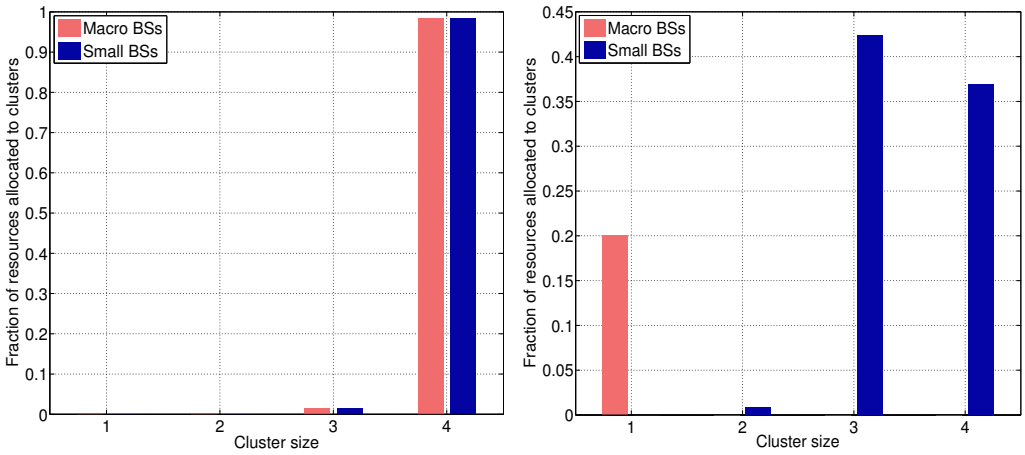
(a) Macro and pico BSs share the resources

(b) Orthogonal resources for macro and pico BSs

Figure 4.8: The number of users served by different clusters, when $\rho = 1$. The "Cluster UEs" refer to the users served by clusters of size larger than 1. Most users have unique association.

cases is very small, which validates our analysis.

We illustrate the fraction of resources allocated to clusters of different sizes at macro and pico BSs in Fig. 4.9. In the case where macro and pico BSs share the resources, we allow clusters to include macro BSs. We observe that most of the BSs transmit in clusters of size 4. On the other hand, in the case where macro and pico BSs use orthogonal resources, only the small BSs are allowd to become a cluster. Most of the small BSs transmit in clusters of sizes 3 and 4, which again implies the potential gain using JT. The resource allocation at BSs with blanking is discussed in the last paragraph when we discuss about Fig. 4.11.



(a) Macro and pico BSs share the resources

(b) Orthogonal resources for macro and pico BSs

Figure 4.9: The fraction of resources allocated by BSs to different clusters, when $\rho = 1$. There are 2 and 3 active ATMs in Figs. 4.9a and 4.9b, respectively.

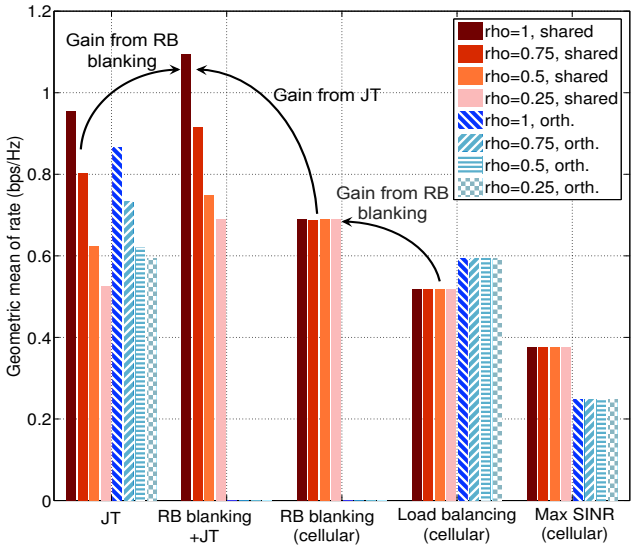


Figure 4.10: The geometric mean of rates using different approaches versus ρ. As ρ decreases, the gain from JT decreases in both shared and orthogonal operation scenarios.

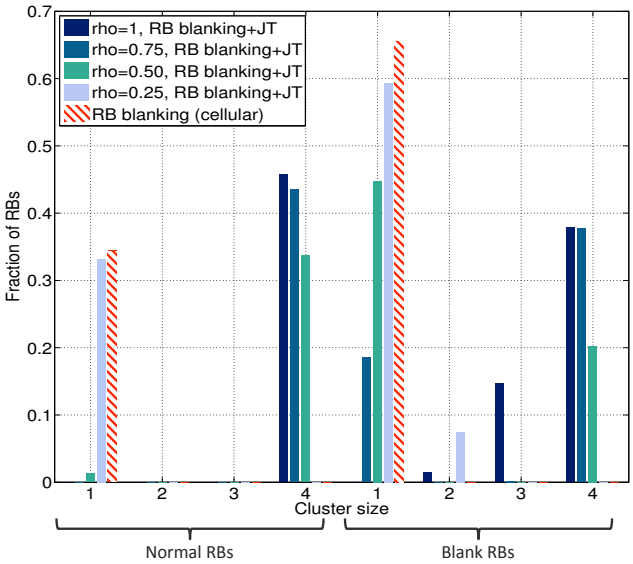


Figure 4.11: The fraction of resources allocated to clusters of different sizes with blanking. As ρ decreases, more resources are allocated to clusters of smaller size.

In Fig. 4.10, we show the geometric mean of rates versus different ρ. We observe that the performance gain using JT decreases as ρ decreases, since the number of users that can be served by clusters decreases. When $\rho \geq 0.5$, the JT can perform better than the cellular transmission. When ρ becomes smaller (e.g., $\rho = 0.25$, where all the clusters we considered can serve the same number of users as in the cellular transmission), the utility using JT is almost the same as the cellular transmission. This implies that the gain from JT increases as more UL pilot resources are available in the system. With limited UL pilot resources, the gain from JT would be quite small.

Fig. 4.11 illustrates the resource allocation for clusters of different sizes versus ρ using RB blanking. The macro BSs are off for about 65% RBs in the

cellular transmission. In the case using JT with RB blanking, as ρ decreases, the clusters serve less users, and more resources are allocated to the clusters of smaller sizes. When $\rho = 0.25$, all resources are allocated to single-BS clusters in normal RBs, and most of the resources are allocated to single-BS clusters in blank RBs. This again suggests that when the available pilot resources are strictly constrained, the gain from JT would be limited.

4.10 Summary

In this chapter, we investigate the joint optimization problem of user association and interference management in massive MIMO HetNets. We consider both the JT and RB blanking approaches for interference management. We first derive the instantaneous rate by exploiting massive MIMO properties, namely the hardening of fast fading to the mean and the independence of peak rate from the user scheduling. We then formulate the joint problem to a convex NUM problem to obtain the desirable user-specific BS clusters and the corresponding resource allocations. The unified formulation can be applied to both JT and blanking approaches, as well as the case where macro and small BSs use orthogonal resources. We further propose an efficient dual subgradient based algorithm, which can be implemented in a partially distributed manner with low overhead. We show this algorithm converges towards the NUM solution. We further show that the NUM solution with the JT approach may not be implementable by a feasible schedule, and thus it provides an upper bound on the performance and can serve as a benchmark. Showing that

most users connect to at most one cluster per RB in heavily loaded networks, we propose to approximate the NUM solution to a unique association, given which we propose a greedy VQ scheduling scheme to provide approximate but implementable results. Simulation results show that the greedy VQ scheduling scheme yields resource allocations that match NUM solutions quite well. More dynamic settings (e.g., users with high mobility) and the study of slot-based algorithm are left for future work. It is also of interest to theoretically bound the gap between the NUM solution and the results of the proposed greedy VQ scheduling scheme.

4.11 Appendix

4.11.1 Proof of Spectral Efficiency Using ZF Precoding

We use $\mathbb{E}\left[\frac{S}{I+N}\right] \approx \frac{\mathbb{E}[S]}{\mathbb{E}[I+N]}$ to approximate the SINR for the calculation of ergodic spectral efficiency in the massive MIMO regime, which is shown to be quite close to the exact asymptotic spectral efficiency [116]. Denoting the kth diagonal element of $\mathbf{A}_j$ by a_{kj} and plugging the precoding matrix $\mathbf{F}_j = \mathbf{G}_j \left(\mathbf{G}_j^H \mathbf{G}_j\right)^{-1} \mathbf{A}_j^{1/2}$ into received signal, the SINR at user k from cluster $\mathcal{C}$ in $\mathcal{A}$ is

$$\mathrm{SINR}_{k\mathcal{C}}^{(\mathcal{A})} = \frac{\sum_{j\in\mathcal{C}} \sum_{l\in\mathcal{C}} \sqrt{\frac{P_j P_l a_{kj} a_{kl}}{S_j(|\mathcal{C}|) S_l(|\mathcal{C}|)}}}{\sigma^2 + \left\| \sum_{l\notin\mathcal{C}} \sum_{u\in\cup_{(\mathcal{C}'\in\mathcal{A}: l\in\mathcal{C}')}} \mathfrak{U}_{\mathcal{C}'}^{(\mathcal{A})}(t) \sqrt{\frac{P_l}{S_l(|\mathcal{C}'|)}} \mathbf{g}_{kl}^H \mathbf{f}_{ul} s_u \right\|^2}. \tag{4.20}$$

Using similar techniques in the proof of Theorem III-1 in [121], we can show $\frac{a_{kj}}{S_j(|\mathcal{C}|)} \to \beta_{kj} \frac{(M_j - S_j(|\mathcal{C}|)+1)}{S_j(|\mathcal{C}|)}$, as $M_j \to \infty$ with fixed ratio $S_j(|\mathcal{C}|)/M_j \leq 1$. Then

we have

$$\sum_{j\in\mathcal{C}}\sum_{l\in\mathcal{C}}\sqrt{\frac{P_jP_la_{kj}a_{kl}}{S_j(|\mathcal{C}|)S_l(|\mathcal{C}|)}}\to\sum_{j\in\mathcal{C}}\sum_{l\in\mathcal{C}}\sqrt{P_jP_l\beta_{kj}\beta_{kl}b_j(|\mathcal{C}|)b_l(|\mathcal{C}|)}, \tag{4.21}$$

where $b_j(|\mathcal{C}|)=\frac{M_j-S_j(|\mathcal{C}|)+1}{S_j(|\mathcal{C}|)}$. As for the interference, we have

$$\begin{aligned}&\left\|\sum_{l\notin\mathcal{C}}\sum_{u\in\cup_{(\mathcal{C}'\in\mathcal{A}:l\in\mathcal{C}')}\mathcal{U}_{\mathcal{C}'}^{(\mathcal{A})}(t)}\sqrt{\frac{P_l}{S_l(|\mathcal{C}'|)}}\mathbf{g}_{kl}^H\mathbf{f}_{ul}s_u\right\|^2\\ =&\sum_{l\notin\mathcal{C}}\sum_{u\in\cup_{(\mathcal{C}'\in\mathcal{A}:l\in\mathcal{C}')}\mathcal{U}_{\mathcal{C}'}^{(\mathcal{A})}(t)}\left\|\sqrt{\frac{P_l}{S_l(|\mathcal{C}'|)}}\mathbf{g}_{kl}^H\mathbf{f}_{ul}\right\|^2\to\sum_{l\notin\mathcal{C}}P_l\beta_{kl},\end{aligned} \tag{4.22}$$

where the last step follows from that channels and precoders of different users are independent. Based on the approximation $\mathbb{E}\left[\frac{S}{I+N}\right]\approx\frac{\mathbb{E}[S]}{\mathbb{E}[I+N]}$, we complete the proof by plugging the above results into (4.20).

4.11.2 Proof of Spectral Efficiency Using MRT Precoding

We first give the following properties of MRT with massive MIMO.

1) We have $\|\mathbf{g}_{kj}\|^2=\mathbf{g}_{jk}^H\mathbf{g}_{kj}=\beta_{kj}\sum_{i=1}^{M_j}h_{kj,i}^*h_{kj,i}$. Recalling that $h_{kj,i}$ are i.i.d. Gaussian, we have $\frac{1}{M_j}\|\mathbf{g}_{kj}\|^2\to\beta_{kj}\mathbb{E}[h_{kj,1}^*h_{jk,1}]=\beta_{kj}$, as M_j and $S_j(|\mathcal{C}|)$ become large with a fixed ratio $S_j(|\mathcal{C}|)/M_j\leq 1$.

2) Plugging $\mathbf{f}_{kj}$, we have $\left|\mathbf{g}_{kj}^H\mathbf{f}_{nj}\right|^2=\left|\mathbf{g}_{kj}^H\frac{\mathbf{g}_{nj}}{\|\mathbf{g}_{nj}\|}\right|^2=\left|\frac{\sqrt{\beta_{kj}\beta_{nj}}}{\|\mathbf{g}_{nj}\|}\sum_{i=1}^{M_j}h_{kj,i}^*h_{jn,i}\right|^2$, which converges to $\frac{\beta_{kj}\beta_{nj}}{\frac{1}{M_j}\|\mathbf{g}_{nj}\|^2}\mathbb{E}\left[|h_{kj,1}^*h_{jn,1}|^2\right]+M_j(M_j-1)\mathbb{E}\left[h_{kj,1}^*h_{nj,1}h_{kj,2}^*h_{nj,2}\right]=\beta_{kj}$ as $M_j\to\infty$, since $h_{kj,i}$ and $h_{nj,i}$ are i.i.d. Gaussian for $n\neq k$.

Using the above two properties and similar techniques as in Appendix

4.11.1, we have

$$\begin{aligned}\mathrm{SINR}_{k\mathcal{C}}^{(\mathcal{A})} \approx & \frac{\sum_{j\in\mathcal{C}}\sum_{l\in\mathcal{C}}\sqrt{\frac{P_j}{S_j(|\mathcal{C}|)}\frac{P_l}{S_l(|\mathcal{C}|)}}\|\mathbf{g}_{kj}\|\|\mathbf{g}_{kl}\|}{\sigma^2+\sum_{j\in\mathcal{C}}(S_j(|\mathcal{C}|)-1)\frac{P_j}{S_j(|\mathcal{C}|)}\beta_{kj}+\sum_{l\notin\mathcal{C}}P_l\beta_{kl}} \\ = & \frac{\sum_{j\in\mathcal{C}}\sum_{l\in\mathcal{C}}\sqrt{\frac{P_jP_lM_jM_l\beta_{kj}\beta_{kl}}{S_j(|\mathcal{C}|)S_l(|\mathcal{C}|)}}}{\sigma^2+\sum_{j\in\mathcal{C}}(S_j(|\mathcal{C}|)-1)\frac{P_j}{S_j(|\mathcal{C}|)}\beta_{kj}+\sum_{l\notin\mathcal{C}}P_l\beta_{kl}}.\end{aligned} \tag{4.23}$$

4.11.3 Proof of Theorem 4.1

If all considered ATSs include clusters with at most one BS, the clusters per ATM can be considered as a new BS. According to Theorem 1 in [83], we can show that there exists at least one feasible schedule that can provide long-term activity fractions approaching the solution of (4.5). For the case where some ATSs including clusters with size larger than 1, we adopt similar techniques in the proof of Theorem 1 in [83]. Details are given as follows.

We denote the set of feasible schedules by $\mathcal{F}$. Let $e_{k\mathcal{C}}^{(\mathcal{A})} \in \{0,1\}$ indicate the instance that user k is served by cluster $\mathcal{C}$, where $e_{k\mathcal{C}}^{(\mathcal{A})}=1$ if user k connects to cluster $\mathcal{C}$ in $\mathcal{A}$ and $e_{k\mathcal{C}}^{(\mathcal{A})}=0$ otherwise. According to Defn. 4.3, the element in $\mathcal{F}$ is a set $\{e_{k\mathcal{C}}^{(\mathcal{A})}\}$ satisfying that user k connects to at most one cluster and BS j serves at most $S_j(L_{\mathcal{A}})$ distinct users in ATM $\mathcal{A}$.

By time sharing among the feasible schedules in $\mathcal{F}$, any fractional association in the convex hull of $\mathcal{F}$ can be achieved in the long term. We denote the convex hull of $\mathcal{F}$ by $X'=\mathrm{conv}(\mathcal{F})$ and the set of feasible activity fractions

by X, i.e.,

$$X = \Bigg\{ x_{k\mathcal{C}}^{(\mathcal{A})} : \sum_{\mathcal{C}: j\in\mathcal{C}, \mathcal{C}\in\mathcal{A}} \sum_{k\in\mathcal{U}} \frac{x_{k\mathcal{C}}^{(\mathcal{A})}}{S_j(L_{\mathcal{A}})} \leq 1, \sum_{\mathcal{C}\in\mathcal{A}} x_{k\mathcal{C}}^{(\mathcal{A})} \leq 1, \\ x_{k\mathcal{C}}^{(\mathcal{A})} \geq 0, \forall k \in \mathcal{U}, \forall j \in \mathcal{B}, \mathcal{C} \in \mathcal{A} \text{ and } \forall \mathcal{A} \Bigg\}.$$

It is easy to show that any feasible schedule in $\mathcal{F}$ satisfies the constraints (4.5b)-(4.5d), and thus $\mathcal{F} \subseteq X$. Note that X is convex. Thus, we have $X' = \text{conv}(\mathcal{F}) \subseteq X$.

As for the opposite direction (i.e., $X \not\subseteq X'$), we first define the *totally unimodular* (TU) matrix: every square submatrix of a TU matrix has determinant $+1$, -1, or 0. The Hoffman & Kruskal's (1956) Theorem claims that a matrix $\mathbf{A}$ is TU if and only if for each integral vector $\mathbf{b}$, the *extreme points* of the polyhedron $\{\mathbf{z} : \mathbf{Az} \leq \mathbf{b}, \mathbf{z} \geq 0\}$ are integral [122]. Let $\mathbf{b} = \left[\mathbf{S}_1^T, \cdots, \mathbf{S}_J^T, 1, \cdots, 1\right]^T$ with $\mathbf{S}_j$ being a $J \times 1$ vector with the same element $S_j(L)$, and $\mathbf{A} = \left[\begin{smallmatrix}\mathbf{B}\\ \mathbf{C}\end{smallmatrix}\right]$ with $\mathbf{B}$ being a diagonal block matrix, whose ith diagonal submatrix is $\mathbf{B}_i = \left[\mathbf{B}_1^{(i)}, \cdots, \mathbf{B}_K^{(i)}\right]$ with the jth row mth column element of $\mathbf{B}_k^{(i)}$ being 1 if $j \in \mathcal{C}_m$ and $|\mathcal{C}_m| = i$, and 0 otherwise. The submatrix $\mathbf{C}$ is a diagonal block matrix with the ith diagonal submatrix $\mathbf{C}_i$ has elements 1 at lth row and kth column if $k \in [|A_i| * (l-1) + 1, |A_i| * l]$ and 0 otherwise, where A_i is the ATM consisted of clusters with size i. The set X can be written in the matrix form as $X = \{\mathbf{x} : \mathbf{Ax} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, where $\mathbf{x} = \left[\mathbf{x}_1^{(1)T}, \cdots, \mathbf{x}_K^{(1)T}, \cdots, \mathbf{x}_1^{(n)T}, \cdots, \mathbf{x}_K^{(n)T}\right]^T$ with $\mathbf{x}_k^{(i)}$ being a vector with the lth element $x_{k\mathcal{C}_l}^{(i)}$. We can show that when there is an ATM consisted of clusters with size larger than 1, the integral matrix $\mathbf{A}$ is not TU, since $\mathbf{A}$ with $|\mathcal{C}| \geq 2$

always includes the submatrix $\left[\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{smallmatrix}\right]$ whose determinant is -2. According to the Hoffman & Kruskal's (1956) Theorem, there are some non-integer extreme points $\mathbf{v} \in X$ that cannot be characterized by a convex combination of any elements in $\mathcal{F}$. Thus, we have $\mathbf{v} \notin \text{conv}(\mathcal{F}) = X'$ and $X \not\subseteq X'$.

4.11.4 Proof of Proposition 4.5

We use the techniques similar to the proof of Prop. 3 in [58], where a graph is used to represent the association, and KKT conditions (4.13) restrict the structure of the graph. For a given ATM $\mathcal{A}$, we denote the graph by G_1, whose nodes represent users, and edge between two nodes represents the BS cluster that serves the two users. Each node has an ID indicating the user index, while each edge has a color that identifies the BS cluster. If there are two users k and m being served by clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ in $\mathcal{A}$ (i.e., $x^{(\mathcal{A})}_{k\mathcal{C}_1} > 0$, $x^{(\mathcal{A})}_{k\mathcal{C}_2} > 0$, $x^{(\mathcal{A})}_{m\mathcal{C}_1} > 0$, $x^{(\mathcal{A})}_{m\mathcal{C}_2} > 0$), we have $R_k = \frac{r^{(\mathcal{A})}_{k\mathcal{C}_1}}{\sum_{j\in\mathcal{C}_1} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_1|)} = \frac{r^{(\mathcal{A})}_{k\mathcal{C}_2}}{\sum_{j\in\mathcal{C}_2} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_2|)}$ and $R_m = \frac{r^{(\mathcal{A})}_{m\mathcal{C}_1}}{\sum_{j\in\mathcal{C}_1} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_1|)} = \frac{r^{(\mathcal{A})}_{m\mathcal{C}_2}}{\sum_{j\in\mathcal{C}_2} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_2|)}$ from KKT condition (4.13), where $R_k = \sum_{\mathcal{A}'} \sum_{\mathcal{C}'\in\mathcal{A}'} x^{(\mathcal{A}')}_{k\mathcal{C}'} r^{(\mathcal{A})}_{k\mathcal{C}'}$. Thus, we have

$$\frac{r^{(\mathcal{A})}_{k\mathcal{C}_1}}{r^{(\mathcal{A})}_{k\mathcal{C}_2}} = \frac{r^{(\mathcal{A})}_{m\mathcal{C}_1}}{r^{(\mathcal{A})}_{m\mathcal{C}_2}}, \tag{4.24}$$

which is true with probability 0. Therefore, it is almost sure that any two users can share at most one same cluster in $\mathcal{A}$. Similarly, we consider an example of three users k, m, i and clusters $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$. We consider the following three cases: 1) If clusters $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are different, we have $\frac{r^{(\mathcal{A})}_{k\mathcal{C}_1}}{r^{(\mathcal{A})}_{k\mathcal{C}_2}} = \frac{\sum_{j\in\mathcal{C}_1} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_1|)}{\sum_{j\in\mathcal{C}_3} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_3|)} \frac{\sum_{j\in\mathcal{C}_3} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_3|)}{\sum_{j\in\mathcal{C}_2} \nu_{j\mathcal{A}}/S_j(|\mathcal{C}_2|)} = \frac{r^{(\mathcal{A})}_{m\mathcal{C}_1}}{r^{(\mathcal{A})}_{m\mathcal{C}_3}} \frac{r^{(\mathcal{A})}_{i\mathcal{C}_3}}{r^{(\mathcal{A})}_{i\mathcal{C}_2}}$, which is true with probability 0.

2) If $\mathcal{C}_1 = \mathcal{C}_2 \neq \mathcal{C}_3$, we have that users m and i are served both by clusters $\mathcal{C}_1$ and $\mathcal{C}_3$, which is true with probability 0 from (4.24).

3) If $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}_3$, we have that users k, m and i are served by the same cluster, which is possible. In this case, the graph becomes a *complete graph*.

Therefore, the graph G_1 with three users either contains a loop with the same color edges or no loop. We can get a similar result for graph G_1 with more than three users, where the users served by the same BS cluster constitute a complete graph. Thus, we generate a new graph, denoted by G_2, where the node represents a cluster. There is an edge between two nodes in G_2, if these two nodes (i.e., clusters) have a common vertex in G_1 (i.e., there is at least one user served by both these two clusters). Thus the number of users who are served by more than one cluster is limited by the edge of G_2. Note that there are $N_{\mathcal{CA}}$ nodes and no loop in G_2. Thus, G_2 is a tree, which has the maximal number of edges being one less than the number of nodes (i.e., $N_{\mathcal{CA}} - 1$).

Chapter 5

Analysis and Optimization of D2D Enhanced Cellular Networks Using Time-Frequency Hopping[1]

Besides small cells, D2D communication is also an emerging offloading approach in HetNets, which reduces energy consumption, efficiently utilizes the network resources, reduces end-to-end latency, and increases the network capacity and flexibility [12–15]. As discussed in Chapter 1, D2D links can either use orthogonal resources or share resources with the cellular network. In a network with orthogonal allocation – called a *dedicated network* – the interference management is simplified, but the resource utilization may be less efficient. On the other hand, if D2D transmissions reuse cellular resources – called a *shared network* – network resources can be used more efficiently, at the cost of a denser interference environment, which complicates the interference management. Which is preferable? Potential D2D data can either be transmitted directly (D2D), or via a base station (BS) – termed *mode selection*. When should a potential D2D link transmit directly, versus relaying via

[1]This chapter has been published in [123]. Coauthor Dr. Mazin Al-Shalash has provided technical suggestions and many insights to this work, and Dr. Constantine Caramanis and Dr. Jeffrey G. Andrews are my supervisors.

the BS? This chapter develops a flexible model to answer these fundamental questions by providing accurate analytical results and simple semi-closed form expressions for performance bounds, which in turn are amenable to efficient optimization. Note that in this dissertation, we focus on the use cases where the D2D traffic is generated by UEs themselves (e.g., sharing a just-taken video), and thus only consider one hop D2D transmission, which is different from the case where D2D works as a relay [124].

5.1 Related Work

Paper [125] investigates both the dedicated and shared approaches for uplink resources and shows that in general, the dedicated approach is more efficient in terms of *transmission capacity*, i.e., it allows more successful transmissions per unit area. On the other hand, in terms of total rate, [126] shows that the shared approach is better in a single cell scenario with a maximal rate cap, taking into account both uplink and downlink transmissions. For a shared network, careful resource allocation is a popular approach to control the mutual interference between cellular and D2D transmissions. For example, an intelligent frequency allocation where orthogonal resources are assigned to nearby cellular and D2D links [127], exclusive D2D transmission zones [128], mixed integer optimization problems [129,130], auction based mechanisms [131,132], a Stackelberg game framework [133], and interference randomization through time hopping [134] are viable approaches for controlling interference. Besides, performance analysis considering interference among D2D links is conducted

in [135]. For mode selection, simple distance-based and received signal-based mode selections are proposed in [136] and [137], respectively. More sophisticated mode selection involving other UEs are proposed in [138–140]. Nevertheless, the majority of earlier studies consider a single cell scenario and propose heuristic algorithms to improve network performance. In this chapter, we leverage tools from stochastic geometry to study a more general D2D-enabled cellular network.

In a pure ad hoc network, there has been significant success over the past decade in proposing tractable models for performance analysis and system design via stochastic geometry [141]. For example, [142] investigates an Aloha-type access mechanism for a large ad hoc network, while [143] analyzes a carrier sense multiple access (CSMA)-type mechanism. The transmission capacity of ad hoc networks are studied and summarized in [144, 145]. There have been some analogous more recent results for cellular networks [146,147], where the BSs are modeled as a Poisson point process (PPP). D2D-enabled cellular networks are essentially a combination of cellular and ad hoc networks, but combining these models into a more general framework is nontrivial. D2D communication can either utilize uplink or downlink resources, and it is not *a priori* clear which resource utilization is better. There has been at least one very recent (parallel) work attempting this for the uplink system [28]. For comparison and the completeness of study, we instead investigate a D2D-enabled cellular network, where downlink resources are either partitioned or shared between D2D and downlink cellular transmissions.

5.2 Contributions

The objective of this chapter is to propose a general framework for the analysis of system performance (e.g., the signal-to-interference-plus-noise ratio (SINR) distribution and total rate) in D2D-enabled cellular networks. We apply this framework to both dedicated and shared downlink networks, which are easy to analyze and optimize, and can be adopted as flexible baseline models for further study. Our key contributions are enabled by simultaneously leveraging techniques from stochastic geometry and optimization theory.

Tractable model for both dedicated and shared cellular networks. We propose a tractable model for a large D2D-enabled cellular network, where the locations of BSs and UEs are modeled as spatial point processes, particularly PPPs. We propose to adopt a time-frequency hopping scheme for potential D2D links to randomize the interference, where each potential D2D link chooses its operation mode (i.e., D2D or cellular mode) at each time slot independently according to a predefined time hopping probability, and accesses each subband independently with a predefined frequency hopping probability. In this model, the derived SINR distributions have remarkably simple forms, which provide an efficient system performance evaluation without time-consuming simulations. It is not always possible to get the expected rate in closed form. We provide a general expression for the average rate and then derive its lower bound, which is in a semi-closed form in interference-limited networks.

Performance optimization and design insights. Based on the

derived SINR distributions and the lower bounds on average rates, we investigate the optimal D2D hopping probabilities (i.e., how often potential D2D links should request a time or frequency slot) using optimization theory. The optimal network performance can serve as a lower bound for D2D-enable cellular networks with more sophisticated scheduling scheme. We find that in many cases, we can either derive the optimal solution in a simple closed-form, or reduce the problem to lower dimension (e.g., one of the hopping probabilities is found in closed-form). The observed design principles are now summarized.

Dedicated vs. shared. Unsurprisingly, the dedicated network has better SINRs since resources allocated to D2D and cellular links are orthogonal. With an optimal spectrum partition between D2D and cellular users, the dedicated network also provides larger average rate, but should be interpreted cautiously. For example, the optimal spectrum partition may be very hard to determine, or it may vary significantly in time or space over a non-homogeneous network (recall we model all BSs and UEs as homogenous PPPs). In such cases, the shared approach may be able to perform significantly better, as it is more flexible. For cases with less local traffic, the shared approach may also have better performance.

Optimal hopping scheme. In the dedicated network with any general non-decreasing utility function, the optimal D2D frequency hopping depends on the service demands of D2D users. D2D links with more traffic to transmit should be more aggressive in their spectrum access, despite the interference that this generates to the rest of the network. The same observation can be

seen from simulation results of the shared network.

As for time hopping, in most considered interference limited cases with heavy load, all potential D2D links should operate in D2D mode (bypassing the BS), assuming the objective is to maximize the total average rate. This result is independent of the average distance between a D2D transmitter and its receiver, which is perhaps surprising, and largely due to the use of total average rate as the utility function. We demonstrate this by giving an example in Section V-A, showing that the optimal mode selection for different utility functions may be very different. In principle, any utility function can be investigated based on the proposed framework, but we use total average rate and leave other utility functions to future work.

5.3 System Model

We focus on a downlink model, where D2D communication uses downlink cellular resources. The key aspects of the model are described in the following subsections.

5.3.1 Deployment of D2D and Cellular Networks

We consider a large D2D-enabled downlink cellular network, illustrated in Fig. 5.1. We classify the potential D2D transmitters into M types which may differ in terms of their service demands and/or the MAC protocol. Note that similar to current wireless traffic growth driven by smartphones proliferating around the world, more local traffic will possibly be generated once the

D2D features are available in future. Therefore, at this stage, the D2D traffic demand as well as its growth is not clear. Though any general distributions can be used to model the location of D2D users, random (uniform) dropping is one of the most popular models in both academia and industry (e.g., [28, 143, 148, 149]). In this chapter, we propose to use the following random dropping model as a first-cut study, and leave other models (e.g., clustered UEs in hotspot) to future work. We assume that the D2D transmitters of the ith type are randomly distributed according to a homogeneous PPP Φ_{D_i} with density λ_{D_i}. The M PPPs are assumed to be independent of each other. Note that the performance of the PPP model can serve as a benchmark for more general settings. Each receiver is assumed to be randomly located around its transmitter according to a two-dimensional Gaussian distribution $N(0, \delta^2)$ with the phase uniformly distributed in $[0, 2\pi]$, so δ parameterizes the distance between the receiver and its transmitter, which is Rayleigh distributed with mean $\delta\sqrt{\frac{\pi}{2}}$ [143]. Other distance distributions can be easily incorporated into the considered framework.

We model the BSs and cellular users as two further independent homogeneous PPPs, denoted by Φ_B and Φ_U with densities λ_B and λ_U, respectively. The model can be easily extended to the case where cellular users have heterogeneous service demands. By tuning the BS, D2D and cellular user densities, along with δ, a very large class of plausible network topologies can be considered with this framework.

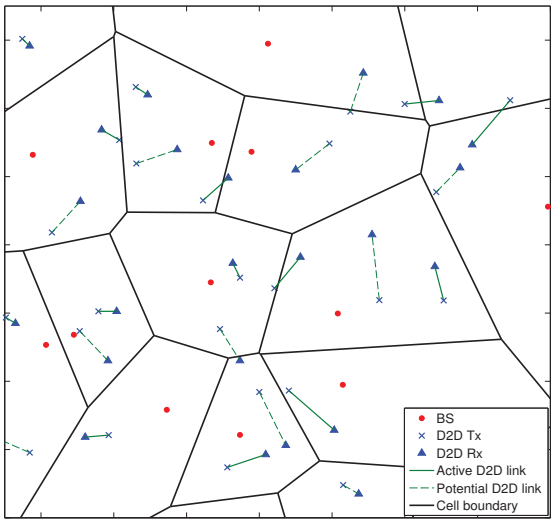


Figure 5.1: Illustration of the network model. The red points are BSs which are deployed according to a PPP. The D2D links include both silent potential D2D links (with dashed lines) and active D2D links (with solid lines).

5.3.2 Scheduling Scheme

We propose to adopt a time-frequency hopping scheme for D2D scheduling, to randomize the occurrence of access collisions with nearby interfering UEs, and thus randomize the strong interference [134]. As illustrated in Fig. 5.2, the time axis is divided into consecutive operation slots. At each slot, the potential D2D links can either be active (i.e., in D2D mode, where traffic is transmitted directly between UEs) or silent (i.e., in cellular mode, where traffic is relayed via the BS), and each potential D2D link selects its operation mode independently. For example, a potential D2D link of type i would either be active with probability $p_{t_i} \in [0, 1]$ or operate in cellular mode with probability $1 - p_{t_i}$. As p_{t_i} increases, more potential D2D links would be in D2D

mode. Thus time hopping is a tool for implementing mode selection, where p_{t_i} results in a tradeoff between spatial reuse and additional interference. In the frequency domain, the ith type D2D links access each subband independently with probability $p_{f_i} \in [0, 1]$. As p_{f_i} increases, more frequency resources are utilized by D2D links at the cost of increasing interference, since more D2D links access the same subbands. Therefore, the frequency hopping probability p_{f_i} results in a tradeoff between frequency efficiency and additional interference.

Using the time-frequency hopping scheme, D2D links are scheduled independently of one another in an Aloha-type fashion in both time and frequency [150]. The outage probability, defined as the probability that the SINR is less than or equal to a given threshold (i.e., $\mathbb{P}(\text{SINR} \leq \beta)$ with β being a predefined threshold), increases almost linearly with the hopping probabilities in the low outage regime, while the spatial reuse increases linearly with the hopping probabilities [144]. So we can adjust the outage probability by changing the hopping probabilities, so as to meet a target outage constraint (i.e., $\mathbb{P}(\text{SINR} \leq \beta) \leq \epsilon$, where ϵ is a predefined parameter). Other scheduling schemes such as centralized approaches or CSMA can be adopted, but with drawbacks in practice (e.g. high overhead) and in terms of tractability (the resulting transmitters are correlated and thus no longer a PPP).

Further, we introduce a penalty assessed to potential D2D links operating in cellular mode, denoted by w, to account for using both uplink and downlink resources. A nominal value for w might be 2, because the local traffic transmitted via a BS requires to establish both the uplink and downlink

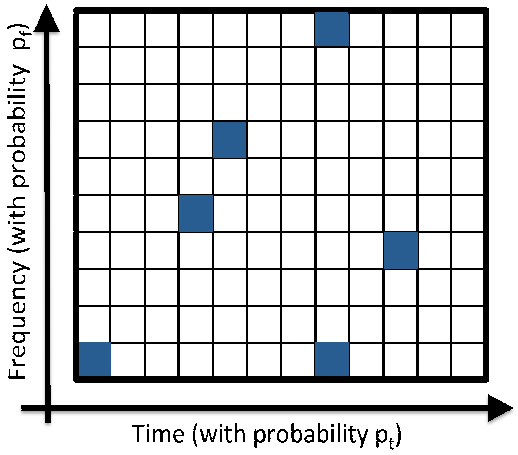


Figure 5.2: Illustration of the time-frequency hopping scheme. The shadowed squares are the RBs occupied by some active D2D links. A potential D2D link accesses each time slot uniformly with probability p_t and accesses each subband uniformly with probability p_f.

transmissions, while the D2D transmission only needs to establish one link. The parameter w can be considered as the price for D2D traffic using cellular mode, which can help adjust the load between D2D and cellular networks.

5.3.3 Load Modeling

We assume that there are B frequency slots (subbands) in the network. We either use the dedicated or the shared spectrum approaches. In a dedicated network, a fraction of resources is allocated to D2D links, denoted by θ, while the rest is allocated to the cellular network. In contrast, in the shared network, the active D2D links share the resources with the cellular network. We assume that the UEs are associated with the nearest BSs, and each BS randomly allocates RBs to its cellular UEs according to their service demands. The

performance using random allocation is a lower bound on more sophisticated scheduling schemes (e.g., ones which are channel or SINR dependent), and the consideration of such schemes is left to future work.

The cellular users are assumed to have the same resource requirement, denoted by b_C. The resource requirement for the ith type D2D links is denoted by b_{D_i}. In this chapter, we consider the resource requirement in terms of the number of subbands for tractability [151]. The ratio of total subbands B to the resource requirement of UEs represents different service demand scenarios (e.g., $b_C = b_{D_i} = B$ for a heavily loaded network). Let N_C and N_{D_i} be the number of original cellular users and of type i potential D2D links in cellular mode, respectively. Let B_C denote the number of available subbands for the cellular network, where $B_C = (1-\theta)B$ in the dedicated network and $B_C = B$ in the shared network. The cell is lightly loaded if $B_C/(b_C N_C + \sum_{i=1}^{M} b_{D_i} N_{D_i}) > 1$, and fully loaded otherwise. In the former case, the BS only transmits on a subset of the subbands, which are called normal RBs, while the other RBs are left blank (i.e., the BS is not transmitting on these RBs). In the latter case, some users have to be blocked (or all UEs are admitted by a cell but can only obtain a fraction of time slots). The admission probability (or fraction of time slots can be obtained) is $B_C/(b_C N_C + \sum_{i=1}^{M} b_{D_i} N_{D_i})$ [151], which is essentially the ratio of the number of available subbands in the cell to the number of subbands needed by cellular users.

5.3.4 Channel Model

In this chapter, we assume that the transmission powers are fixed at P_D for D2D transmitters and P_B for BSs. General attenuation functions can be adopted, but we focus on the standard power law attenuation function $l(d) = d^{-\alpha}$, where d is the distance from the transmitter to a receiver, and α is the path loss exponent. We assume all links experience independent Rayleigh fading. Shadowing is not explicitly modeled, but is already captured by the randomness of PPP in some sense, e.g., [152] showed that a grid BS model with fairly strong (standard deviation greater than 10dB) log-normal shadowing is nearly equivalent to a PPP model without shadowing.

The notations are summarized in Table 5.1.

5.4 Analysis of the Dedicated Network

In this section, we investigate the key performance metrics in the dedicated network.

5.4.1 SINR Distribution

Due to the stationarity of Φ_D, we conduct analysis on a typical D2D receiver located at the origin, whose transmitter is active. The location of the typical transmitter is denoted by X_0. For simplicity, we denote D2D links by the location of their transmitters (e.g., the typical link is called link X_0). According to the load model described in Section 5.3.3, the network has blank RBs when it is under-loaded. To get the average fraction of blank RBs, we

Table 5.1: Notation summary for D2D-cellular networks with time-frequency hopping

Notation	Description
Φ_{D_i}	PPP of type i D2D links
Φ_B,Φ_U	PPP of BSs, cellular UEs
λ_{D_i}	Density of type i D2D links
λ_B, λ_U	Density of BSs, cellular UEs
M	Number of D2D types
δ	Parameter of distance between D2D transmitter and its receiver
p_{f_i}	Frequency hopping probability
p_{t_i}	Time hopping probability
w	Penalty for potential D2D links in cellular mode
θ	Fraction of resource allocated to D2D
B	Total frequency subbands
b_{D_i}	Service demand of type i D2D links
b_C	Service demand of cellular users
P_B	Transmit power of BSs
P_D	Transmit power of D2D transmitters
σ^2	Noise power
α	Path loss exponent
$\rho^{(O)}$, $\rho^{(S)}$	Fraction of normal RBs in the dedicated and shared network, respectively
$p_a^{(O)}$, $p_a^{(S)}$	Admission probability in the dedicated and shared network, respectively
$P_D^{(O)}$, $P_D^{(S)}$	Coverage probability of D2D links in the dedicated and shared network, respectively
$P_C^{(O)}$, $P_C^{(S)}$	Coverage probability of cellular UEs in the dedicated and shared network, respectively
$R_D^{(O)}$, $R_D^{(S)}$	Rate of D2D links in the dedicated and shared network, respectively
$R_C^{(O)}$, $R_C^{(S)}$	Rate of cellular UEs in the dedicated and shared network, respectively

first find the average load per cell. The average coverage area of a BS is $\frac{1}{\lambda_B}$ [153]. Thus, the average numbers of cellular users and of ith-type D2D links in cellular mode in a cell are λ_U/λ_B and $(1-p_{t_i})\lambda_{D_i}/\lambda_B$, respectively. Recalling the available fraction of subbands for the cellular network is $(1-\theta)B$, the average fraction of normal RBs sent by a BS is approximated by

$$\rho^{(O)} \approx \min\{\frac{b_C\lambda_U + \sum_{i=1}^{M}(1-p_{t_i})b_{D_i}\lambda_{D_i}}{\lambda_B(1-\theta)B}, 1\}. \tag{5.1}$$

Assuming that BSs randomly allocate RBs to cellular UEs, the set of active interfering BSs at a typical RB in the dedicated network, denoted by $\tilde{\Phi}_B$, can be considered as a thinning process from the baseline BS process Φ_B, which is approximated by a PPP with density $\rho^{(O)}\lambda_B$. This approximation is validated in Section 5.7, where we observe that analysis and simulation results are in good agreement.

Adopting time-frequency hopping, the interfering D2D transmitters are those which access the same time-frequency RBs. We denote the set of type i interfering D2D transmitters by $\tilde{\Phi}_{D_i}$, which is a thinning process from Φ_{D_i}. Based on the Thinning Theorem of PPP [153], the thinning process $\tilde{\Phi}_{D_i}$ is a PPP with density $p_{t_i}p_{f_i}\lambda_{D_i}$. Applying the superposition of PPPs [153], the set of interfering transmitters can be considered as a single PPP $\tilde{\Phi}_D$ with density $\tilde{\lambda}_D = \sum_{i=1}^{M} p_{t_i}p_{f_i}\lambda_{D_i}$.

The SINR of the typical D2D link is

$$\text{SINR} = \frac{P_D h_0 |X_0|^{-\alpha}}{I_{\tilde{\Phi}_D} + \sigma^2},$$

where $I_{\tilde{\Phi}_D} = \sum_{X_i \in \tilde{\Phi}_D \setminus X_0} P_D h_i |X_i|^{-\alpha}$ is the interference from other D2D users, and σ^2 is the noise power. The SINR complementary cumulative distribution function (CCDF) of the D2D links, also known as the coverage probability, is given by Proposition 5.1.

Proposition 5.1. *The SINR CCDF of D2D links in the dedicated network is*

$$\begin{aligned} &\mathbb{P}_D^{(O)}(\beta) \triangleq \mathbb{P}\left(SINR > \beta\right) \\ &= \int_0^\infty e^{-\beta P_D^{-1} \sigma^2 v^\alpha} \mathcal{L}_{I_{\tilde{\Phi}_D}}(\beta P_D^{-1} v^\alpha) \frac{v e^{-\frac{v^2}{2\delta^2}}}{\delta^2} dv. \end{aligned} \tag{5.2}$$

where the Laplace transform of interference from D2D links is

$$\mathcal{L}_{I_{\tilde{\Phi}_D}}(s) = \exp\left(-\tilde{\lambda}_D \frac{2\pi^2/\alpha}{\sin\left(2\pi/\alpha\right)} (sP_D)^{\frac{2}{\alpha}}\right), \tag{5.3}$$

Proof. See Appendix 5.9.1. □

The above SINR CCDF indicates that the time and frequency hopping impact the SINR of D2D links in a product term $p_{t_i} p_{f_i}$ in $\tilde{\lambda}_D$. That is, as long as the product $p_{t_i} p_{f_i}$ is a constant, the network performance will stay the same. There is a tradeoff between the density and the performance of D2D links: as more and more potential D2D transmitters attempt to transmit, though the density of active links increases, the interference increases and thus the SINR decreases.

Assuming the nearest-BS association, the probability density function of the distance between a user and its associated BS is $f_r(r) = e^{-\lambda_B \pi r^2} 2\lambda_B \pi r$ [154]. In the dedicated network, there is no D2D-cellular interference. Therefore, we can leverage the analytical results of the cellular network in [146] to

evaluate the cellular network performance, where the SINR CCDF of cellular users is given by Proposition 5.2.

Proposition 5.2. *The SINR CCDF of a typical cellular user in the dedicated network is*

$$\mathbb{P}_C^{(O)}(\beta) \triangleq \mathbb{P}(SINR > \beta) = \int_0^\infty e^{-s\sigma^2} \mathcal{L}_{I_{\tilde{\Phi}_B}}(s) f_r(r) dr. \tag{5.4}$$

where $s = \beta P_B^{-1} r^\alpha$*, and the Laplace transform of interference from cellular UEs is*

$$\mathcal{L}_{I_{\tilde{\Phi}_B}}(\beta P_B^{-1} r^\alpha) = \exp\left(-2\pi\rho^{(O)}\lambda_B r^2 H_1(\beta,\alpha)\right), \tag{5.5}$$

with $H_1(\beta,\alpha) = \int_1^\infty \frac{x}{1+\beta^{-1}x^\alpha} dx$*.*

Proof. Proof is given by Appendix B in [146], where we simplify the result by letting $x = v/r$ in (21) of [146]. □

When the network is interference-limited (i.e., the thermal noise is ignored), the above results can be further simplified:

Corollary 5.1. *When* $\sigma^2 \to 0$*, the SINR CCDFs of D2D links and cellular users, respectively, are*

$$\mathbb{P}_D^{(O)}(\beta) = \frac{1}{1 + 2\delta^2 \tilde{\lambda}_D \frac{2\pi^2/\alpha}{\sin(2\pi/\alpha)} \beta^{\frac{2}{\alpha}}}, \tag{5.6}$$

$$\mathbb{P}_C^{(O)}(\beta) = \frac{1}{2\rho^{(O)} H_1(\beta,\alpha) + 1}. \tag{5.7}$$

5.4.2 Rate Analysis in the Dedicated Network

In this section, we analyze the average rates of cellular and D2D links in the dedicated network. By treating the interference as noise, we use Shannon's capacity formula to approximate the rate, i.e., $W \log_2(1 + \text{SINR})$, where W is the available bandwidth. Assuming the fraction of time slots is T, the long-term rate becomes $R = TW \log_2(1 + \text{SINR})$, where TW can be considered as the total available fraction of RBs.

Recall that the admission probability (i.e. available fraction of time slots) of cellular UEs is

$$p_a^{(O)} = \min\left\{\frac{(1-\theta)B}{b_C\bar{N}_C + \sum_{i=1}^{M} b_{D_i}\bar{N}_{D_i}},\ 1\right\},$$

where $\bar{N}_C$ and $\bar{N}_{D_i}$ are the expected number of cellular UEs and of type i potential D2D links in cellular mode in the typical user associated cell, respectively. Note that a random UE is more likely to connect to a cell with larger coverage area. Denoting the BS serving the typical UE by B_0, the expected coverage area of BS B_0 is larger than $1/\lambda_B$, known as Feller's paradox [155]. The average coverage area of BS B_0 is instead given by $9/(7\lambda_B)$ [79, 156]. Therefore, similar to [151], the admission probability can be approximated to

$$p_a^{(O)} \approx \min\left\{\frac{7(1-\theta)B\lambda_B}{9\left(b_C\lambda_U + \sum_{i=1}^{M} b_{D_i}(1-p_{t_i})\lambda_{D_i}\right)},\ 1\right\}. \tag{5.8}$$

Recalling that w is the price for a D2D link operating in cellular mode, the average rates of cellular users and D2D links are given by Theorem 5.1.

Theorem 5.1. *The average achievable rates of a typical cellular user and a D2D link of the ith type, respectively, are*

$$R_C^{(O)} = b_C p_a^{(O)} \int_0^\infty \frac{\log_2(e)}{(\beta+1)} \mathbb{P}_C^{(O)}(\beta)\, d\beta, \tag{5.9}$$

$$\begin{aligned} R_{D_i}^{(O)} =& \min\{p_{f_i}\theta B, b_{D_i}\} p_{t_i} \int_0^\infty \frac{\log_2(e)}{(\beta+1)} \mathbb{P}_D^{(O)}(\beta)\, d\beta \\ &+ \frac{b_{D_i}}{w}(1-p_{t_i}) p_a^{(O)} \int_0^\infty \frac{\log_2(e)}{(\beta+1)} \mathbb{P}_C^{(O)}(\beta)\, d\beta, \end{aligned} \tag{5.10}$$

where $\mathbb{P}_C^{(O)}(\beta)$ *and* $\mathbb{P}_D^{(O)}(\beta)$ *are given by (5.4) and (5.2), respectively. Further, we can get the average rate of a typical D2D link by* $R_D^{(O)} = \sum_{i=1}^M (\lambda_{D_i}/\lambda_D) R_{D_i}^{(O)}$.

Proof. The long term rate of a typical cellular user is

$$\begin{aligned} R_C^{(O)} =& b_C p_a^{(O)} \mathbb{E}\left[\log_2(1+\mathrm{SINR})\right] \\ =& b_C p_a^{(O)} \int_0^\infty \mathbb{P}\left(\mathrm{SINR} > 2^t - 1\right) dt \\ =& b_C p_a^{(O)} \int_0^\infty \frac{\log_2(e)}{(\beta+1)} \mathbb{P}(\mathrm{SINR} > \beta) d\beta \end{aligned} \tag{5.11}$$

where we let $2^t - 1 = \beta$ in the last equality. A D2D link can be either in D2D mode or cellular mode, and thus the average rate of a type i D2D link is

$$R_{D_i}^{(O)} = \mathbb{P}(\text{D2D mode}) \mathbb{E}\left[R_{\text{D2D mode}}^{(O)}\right] + \mathbb{P}(\text{cellular mode}) \frac{1}{w} \mathbb{E}\left[R_{\text{cellular mode}}^{(O)}\right],$$

where the rate obtained in D2D mode is

$$\mathbb{E}\left[R_{\text{D2D mode}}^{(O)}\right] = \min\{p_{f_i}\theta B, b_{D_i}\} \int_0^\infty \frac{\log_2(e)}{(\beta+1)} \mathbb{P}_D^{(O)}(\beta)\, d\beta, \tag{5.12}$$

and the rate obtained in cellular mode is approximated to the rate in the downlink system for tractability:

$$\begin{aligned}&\mathbb{E}\left[R^{(O)}_{\text{cellular mode}}\right]\\ =&\min\left\{\mathbb{E}\left[R^{(O)}_{\text{cellular mode in DL}}\right],\mathbb{E}\left[R^{(O)}_{\text{cellular mode in UL}}\right]\right\}\\ \approx&\mathbb{E}\left[R^{(O)}_{\text{cellular mode in DL}}\right]=\frac{b_{D_i}}{b_C}R^{(O)}_C.\end{aligned}$$

Note that in the rate derivation in this chapter, we assume that the number of users associated with the BS serving the typical link and the SINR distribution of the typical link are independent, and thus plugging (5.2) and (5.4) into (5.12) and (5.11), respectively, the proof is complete. □

5.5 Analysis of the Shared Network

In this section, we turn our attention to the shared network, where the resources are reused between D2D and cellular networks, and thus there is D2D-cellular interference.

5.5.1 SINR Distribution of D2D links

As in Section 5.4, let $\tilde{\Phi}_D$ be the set of interfering D2D links, which is a PPP with density $\tilde{\lambda}_D$. The average fraction of normal RBs in the shared network is approximated by

$$\rho^{(S)}\approx\min\{\frac{b_C\lambda_U+\sum_{i=1}^{M}b_{D_i}(1-p_{t_i})\lambda_{D_i}}{\lambda_B B},1\}.$$

We again consider a typical active D2D receiver located at the origin. Taking into account now the interference from both cellular and D2D networks,

the D2D SINR is

$$\mathrm{SINR}_D = \frac{P_D h_0 |X_0|^{-\alpha}}{I_{\tilde{\Phi}_D} + I_{\tilde{\Phi}_B} + \sigma^2},$$

where X_0 is the location of the typical D2D transmitter, the interference from other D2D transmitters is $I_{\tilde{\Phi}_D} = \sum_{X_i \in \tilde{\Phi}_D \setminus X_0} P_D h_i |X_i|^{-\alpha}$, and the interference from BSs is $I_{\tilde{\Phi}_B} = \sum_{B_i \in \tilde{\Phi}_B} P_B h_i |B_i|^{-\alpha}$.

Theorem 5.2. *The SINR distribution of an active D2D link in the shared network is*

$$\begin{aligned} \mathbb{P}_D^{(S)}(\beta) &\triangleq \mathbb{P}\left(SINR_D > \beta\right) \\ &= \int_0^\infty e^{-s\sigma^2} \mathcal{L}_{I_{\tilde{\Phi}_D}}(s) \mathcal{L}_{I_{\tilde{\Phi}_B}}(s) \frac{v}{\delta^2} \exp\left(-\frac{v^2}{2\delta^2}\right) dv, \end{aligned} \tag{5.13}$$

where $s = \beta P_D^{-1} v^\alpha$, $\mathcal{L}_{I_{\tilde{\Phi}_D}}(s)$ *can be calculated according to (5.3), and the Laplace transform of interference from BSs is*

$$\mathcal{L}_{I_{\tilde{\Phi}_B}}(s) = \exp\left(-2\pi\rho^{(S)}\lambda_B v^2 H_0(\beta, \alpha)\right), \tag{5.14}$$

with $H_0(\beta, \alpha) = \int_0^\infty \frac{x}{1+\beta^{-1}P_D/P_B x^\alpha} dx$.

Proof. Given the distance from D2D transmitter to its receiver, denoted by v, the conditional coverage probability is

$$\mathbb{P}(\mathrm{SINR}(v) > \beta \mid v) = \exp(-s\sigma^2) \mathcal{L}_{I_{\tilde{\Phi}_D}}(s) \mathcal{L}_{I_{\tilde{\Phi}_B}}(s),$$

due to that h_0 is Rayleigh fading and $I_{\tilde{\Phi}_D}$ is independent of $I_{\tilde{\Phi}_B}$. The $\mathcal{L}_{I_{\tilde{\Phi}_D}}(s)$

is given by (5.3). Similarly, we have

$$\begin{aligned}\mathcal{L}_{I_{\tilde{\Phi}_B}}(s) =& \mathbb{E}\left[\prod_{Z_i\in\tilde{\Phi}_B}\frac{1}{1+sP_B|B_i|^{-\alpha}}\right]\\ =& \exp\left(\int_0^\infty \frac{-2\pi\rho^{(S)}\lambda_B r}{1+\beta^{-1}P_D/P_B(r/v)^\alpha}dr\right)\\ =& \exp\left(\int_0^\infty \frac{-2\pi\rho^{(S)}\lambda_B v^2 x}{1+\beta^{-1}P_D/P_B x^\alpha}dx\right),\end{aligned}$$

where the last equality is obtained by letting $x = r/v$. Letting $H_0(\beta,\alpha) = \int_0^\infty \frac{x}{1+\beta^{-1}P_D/P_B x^\alpha}dx$, the proof is complete. □

Theorem 5.2 shows that for any given SINR threshold, the coverage probability of D2D links is monotonically decreasing as the access probabilities p_{t_i} and/or p_{f_i} increase, due to the increasing interference from the D2D network. On the other hand, the relationship between the MAC protocol and average rate is more subtle, and is discussed in Section 5.5.3.

5.5.2 SINR Distribution of Cellular Users

In this section, we conduct an analysis on a typical cellular UE, which is assumed to be located at the origin.

Theorem 5.3. *The SINR distribution of a typical cellular user in the shared network is*

$$\begin{aligned}&\mathbb{P}_C^{(S)}(\beta) \triangleq \mathbb{P}\left(SINR > \beta\right)\\ =&\int_0^\infty e^{-s\sigma^2}\mathcal{L}_{I_{\tilde{\Phi}_D}}(s)\mathcal{L}_{I_{\tilde{\Phi}_B}}(s)e^{-\lambda_B\pi r^2}2\lambda_B\pi r dr,\end{aligned} \tag{5.15}$$

where $s = \beta P_B^{-1} r^{\alpha}$*, and the Laplace transform of interference from D2D links* $\mathcal{L}_{I_{\tilde{\Phi}_D}}(s)$ *is given by (5.3). The Laplace transform of cellular interference* $\mathcal{L}_{I_{\tilde{\Phi}_B}}(s)$ *can be obtained by (5.5), where* $\rho^{(O)}$ *should be replaced by* $\rho^{(S)}$*.*

Proof. We omit the proof as it is similar to the proof of Theorem 5.2. □

Corollary 5.2. *When* $\sigma^2 \to 0$*, the SINR CCDF of D2D links and of cellular users are, respectively,*

$$\mathbb{P}_D^{(S)}(\beta) = \frac{1}{2\delta^2 \tilde{\lambda}_D \kappa \pi \beta^{\frac{2}{\alpha}} + 4\delta^2 \pi \rho^{(S)} \lambda_B H_0(\beta, \alpha) + 1}, \tag{5.16}$$

$$\mathbb{P}_C^{(S)}(\beta) = \frac{1}{\frac{\tilde{\lambda}_D}{\lambda_B} \kappa (\beta \frac{P_D}{P_B})^{\frac{2}{\alpha}} + 2\rho^{(S)} H_1(\beta, \alpha) + 1}, \tag{5.17}$$

where $\kappa = \frac{2\pi/\alpha}{\sin(2\pi/\alpha)}$*.*

Proof. See Appendix 5.9.2. □

From the above analysis, we can see that the coverage probabilities of D2D links and of cellular UEs are both decreasing functions of p_{t_i} and p_{f_i}, due to the increasing interference as more D2D links access the same RBs.

5.5.3 Rate Analysis in the Shared Network

Similar to the dedicated system, the admission probability of cellular users is

$$p_a^{(S)} \approx \min \left\{ \frac{7B\lambda_B}{9\left(b_C \lambda_U + \sum_{i=1}^{M} b_{D_i}(1 - p_{t_i})\lambda_{D_i}\right)},\ 1 \right\}. \tag{5.18}$$

The average rates of cellular users and D2D links in the shared network are given by Proposition 5.3.

Proposition 5.3. *The average achievable rates of a cellular user and of a type i D2D link, respectively, are*

$$R_C^{(S)} = b_C p_a^{(S)} \int_0^\infty \frac{\log_2(e)}{(\beta+1)} \mathbb{P}_C^{(S)}(\beta)\, d\beta, \tag{5.19}$$

$$\begin{aligned} R_{D_i}^{(S)} = {} & p_{t_i} \min\{p_{f_i} B, b_{D_i}\} \int_0^\infty \frac{\log_2(e)}{(\beta+1)} \mathbb{P}_D^{(S)}(\beta)\, d\beta \\ & + \frac{b_{D_i}}{b_C w}(1-p_{t_i}) R_C^{(S)}, \end{aligned} \tag{5.20}$$

where $\mathbb{P}_C^{(S)}(\beta)$ *is given by (5.15) and* $\mathbb{P}_D^{(S)}(\beta)$ *is given by (5.13).*

Then we can get the average rate of a typical D2D link as $R_D^{(S)} = \sum_{i=1}^M \frac{\lambda_{D_i}}{\lambda_D} R_{D_i}^{(S)}$.

Corollary 5.3. *We further have lower bounds on the rates as*

$$R_C^{(S)} \geq R_{Cl}^{(S)} = \sup_\beta \; b_C p_a^{(S)} \log_2(1+\beta) \mathbb{P}_C^{(S)}(\beta), \tag{5.21}$$

and $R_{D_i}^{(S)} \geq R_{Dl_i}^{(S)}$*, where*

$$R_{Dl_i}^{(S)} = \sup_\beta \left(p_{t_i} \min\{p_{f_i} B, b_{D_i}\} \log_2(\beta+1) \mathbb{P}_D^{(S)}(\beta) \right) + \frac{b_{D_j}}{b_C w}(1-p_{t_i}) R_{Cl}^{(S)}. \tag{5.22}$$

Proof. Denoting $\Gamma = \text{SINR}$, for any β, we have

$$\begin{aligned} \mathbb{E}\left[\log_2(1+\Gamma)\right] = {} & \mathbb{P}(\Gamma > \beta) \mathbb{E}\left[\log_2(1+\Gamma) \mid \Gamma > \beta\right] \\ & + \mathbb{P}(\Gamma \leq \beta) \mathbb{E}\left[\log_2(1+\Gamma) \mid \Gamma \leq \beta\right] \\ \geq {} & \mathbb{P}(\Gamma > \beta) \mathbb{E}\left[\log_2(1+\Gamma) \mid \Gamma > \beta\right] \\ \geq {} & \mathbb{P}(\Gamma > \beta) \log_2(1+\beta), \end{aligned}$$

and thus $\mathbb{E}\left[\log_2(1+\Gamma)\right] \geq \sup_\beta \mathbb{P}(\Gamma > \beta) \log_2(1+\beta)$. □

The above lower bounds can also be extended to the dedicated network.

According to the above analysis, when the frequency hopping probability p_{f_i} increases, the average rate of cellular users decreases, because the interference from D2D links increases. Thus the cellular rate is a monotonic function of p_{f_i}. As for the time hopping, when p_{t_i} increases, more potential D2D links would operate in D2D mode. On the one hand, the interference from D2D links increases as p_{t_i} increases, leading to a lower SINR of the cellular links; on the other hand, the cellular links benefit from D2D offloading, since more resources would be available for the remaining cellular links. Therefore, it is difficult to determine the impact of time hopping on the rate of cellular users. As for the rate of D2D links, it is even more difficult to explore the impact of time-frequency hopping, because both time and frequency hopping result in the tradeoff between resource efficiency and additional interference. It is not *a priori* clear whether larger time and frequency hopping probabilities would be beneficial or not. However, by changing variables, we can get the optimal solution of at least one variable and thus reduce the dimensions of the optimization problem. We explore these issues in detail in the next section.

5.6 Optimization of the D2D-Cellular Network

Based on the above analytical results, we now turn our attention to the optimization of network performance. As in [125,144,153], we study the utility maximization in the interference limited network (i.e., $\sigma^2 \to 0$) for simplicity.

5.6.1 Optimization of the Dedicated Network

The utility functions of a cellular user and of a type i D2D link are denoted by $U_C(R_C)$ and $U_D(R_{D_i})$, respectively, where $U_D(\cdot)$ and $U_C(\cdot)$ are continuously differentiable, non-decreasing, and concave functions [49]. The optimization problem can be formulated as

$$\begin{aligned} \max_{p_t, p_f} \quad & \sum_{i=1}^{M} \lambda_{D_i} U_D(R_{D_i}^{(O)}) + \lambda_U U_C(R_C^{(O)}) \\ \text{s.t.} \quad & 0 \le p_{t_i} \le 1, 0 \le p_{f_i} \le 1, \end{aligned} \tag{5.23}$$

where $R_{D_i}^{(O)}$ and $R_C^{(O)}$ are given by Theorem 5.1.

For the optimal frequency hopping probability, we obtain closed-form solution that is *independent of the choice of utility functions*.

Proposition 5.4. *For any non-decreasing utility function, the optimal frequency hopping probability in the dedicated network is* $p_{f_i}^* = \min\{1, b_{D_i}/(\theta B)\}$.

Proof. The objective function is a non-decreasing function of p_{f_i} when $p_{f_i}\theta B \le b_{D_i}$, and becomes monotonically decreasing when $p_{f_i}\theta B > b_{D_i}$. Therefore, the optimal frequency hopping probability is $p_{f_i}^* = \min\{1, b_{D_i}/(\theta B)\}$. □

The above proposition shows that the D2D network is resource limited. The larger the service demand is, the more aggressive the D2D link should be to access the frequency bands.

Though sum rate maximization may not be a good performance metric in the sense that it has not considered fairness among UEs, it is a reasonable objective function for a first-cut investigation of the complicated hybrid

network. Therefore, in the following, we focus on the linear utility function $U(x) = x$. A single tier cellular network is heavily loaded in most cases (e.g., in a typical LTE network with $B_C = 10$MHz, $b_C \approx 1$MHz, and $\frac{\lambda_U}{\lambda_B} > 10$, we have $B_C\lambda_B < b_C\lambda_U$). Therefore, we consider a congested network where $7B_C\lambda_B < 9b_C\lambda_U$, and thus $\rho^{(O)} = 1$ (i.e., BSs always send normal RBs) and $p_a^{(O)} = \frac{7B_C\lambda_B}{9\left(b_C\lambda_U + \sum_{i=1}^{M} b_{D_i}(1-p_{t_i})\lambda_{D_i}\right)}$.

Definition 5.1. *The rate density is defined as the expected total rate of D2D links and cellular users per surface unit.*

For tractability, we investigate the hopping scheme to maximize the rate lower bounds given by Corollary 5.3. We compare the results of exact rates and their lower bounds by simulations in Section 5.7. Note that the following results can be easily extended to the cases where the Modulation and Coding Scheme (MCS) is not adaptive by setting a fixed β. Using the rate lower bounds, the rate density of dedicated and shared networks can be respectively calculated by

$$d_{\text{rate}}^{(O)} \triangleq \sum_{j=1}^{M} \lambda_{D_j} R_{Dl_j}^{(O)} + \lambda_U R_{Cl}^{(O)}, \tag{5.24}$$

and

$$d_{\text{rate}}^{(S)} = \sum_{j=1}^{M} \lambda_{D_j} R_{Dl_j}^{(S)} + \lambda_U R_{Cl}^{(S)}, \tag{5.25}$$

where $R_{Dl_j}^{(O)}$, $R_{Cl}^{(O)}$, $R_{Dl_j}^{(S)}$ and $R_{Cl}^{(S)}$ are given by Corollary 5.3.

Proposition 5.5. *To maximize the rate density in the dedicated network with $w \geq 1$, we have $p_{t_i}^* = 1, \forall i = \{1, \cdots, M\}$. On the other hand, when $w \to 0$, we have $p_{t_i}^* \to 0$.*

Proof. See Appendix 5.9.3. □

Propositions 5.4 and 5.5 imply that both D2D and cellular networks are resource limited when the network is fully loaded. In order to utilize resources efficiently, all potential D2D links should be in D2D mode when $w \geq 1$. On the other hand, by setting w small enough, the potential D2D links can be pushed to cellular mode (i.e., $p_{t_i}^* \to 0$).

Note that analytically it is true that all potential D2D links are in D2D mode to maximize the total average rate. However, traffic channels in real cellular systems are typically not designed to operate at very low SINR (e.g., SINR < -6dB) [157]. If the average distance is very large such that the SINRs of many D2D links are smaller than -6dB, the optimal mode selection would be different. Also, maximization of different utility functions would lead to different optimal mode selections. For example, when we consider the max-min utility, the optimal time hopping would depend on the average D2D link length, which is characterized by δ. As δ increases, the rates of potential D2D links obtained in D2D mode decrease, and may be smaller than the rate obtained in cellular mode. Thus with an increasing probability, the potential D2D links in D2D mode would have the minimal rates in the system. Therefore, we would push some potential D2D links to cellular mode in this example, in order to increase their rates and maximize the minimal rate (i.e., optimal time hopping probability $p_t^* < 1$).

Given the optimal time hopping and frequency hopping, we investi-

gate the optimal resource partition between D2D and cellular networks (i.e., θ). Plugging $p_{f_i}^* = \min\{1, \frac{b_{D_i}}{\theta B}\}$ and $p_{t_i}^* = 1$ to (5.24), the objective function is a non-differentiable function of θ. We denote $\tilde{b}_i = b_{D_i}/B$ for $i = 1, \cdots, M$ and $\tilde{b}_0 = 0$ for simplicity. Without loss of generality, we assume the sequence $\{\tilde{b}_i\}_{i=0}^M$ is in ascending order (i.e., $\tilde{b}_0$ is the smallest and $\tilde{b}_M$ is the largest). Let $\tilde{b}_L$ be the largest $\tilde{b}_i$ that is smaller than 1. We partition the domain of θ into $\left[\tilde{b}_i, \min\left\{\tilde{b}_{i+1}, 1\right\}\right], i = 0, \cdots, L$. On the ith region $\left[\tilde{b}_i, \min\left\{\tilde{b}_{i+1}, 1\right\}\right]$, the types of D2D links can be separated into two sets, where $\mathcal{S}_i = \{0, \cdots, i\}$ and $\mathcal{G}_i = \{i+1, \cdots, M\}$. We have $p_{f_j}^* = \frac{\tilde{b}_j}{\theta}$ for $j \in \mathcal{S}_i$, and $p_{f_j}^* = 1$ for $j \in \mathcal{G}_i$. Thus, the objective function becomes a differentiable function on each partition. Denote $A_i = B\log_2(\beta_D + 1)\sum_{j\in\mathcal{S}_i}\lambda_{D_j}\tilde{b}_j$, $C_i = 2\delta^2\frac{2\pi^2/\alpha}{\sin(2\pi/\alpha)}\beta^{\frac{2}{\alpha}}\sum_{j\in\mathcal{S}_i}\lambda_{D_j}\tilde{b}_j$, $D = \frac{7B\lambda_B}{9}\frac{\log_2(\beta_C+1)}{2H_1(\beta_C,\alpha)+1}$, $E_i = B\log_2(\beta_D + 1)\sum_{j\in\mathcal{G}_i}\lambda_{D_j}$, and $F_i = 2\delta^2\sum_{j\in\mathcal{G}_i}\lambda_{D_j}\frac{2\pi^2/\alpha}{\sin(2\pi/\alpha)}\beta^{\frac{2}{\alpha}} + 1$. Letting

$$\tilde{b}_i' = \begin{cases} 1, \text{ if } E_i > DF_i, \\ \sqrt{\frac{C_i(A_iF_i - E_iC_i)}{F_i^2(DF_i - E_i}}) - \frac{C_i}{F_i}, \text{ otherwise,} \end{cases} \tag{5.26}$$

we can express the optimal solution θ^* in terms of $\tilde{b}_i'$ as follows.

Proposition 5.6. *The optimal θ to maximize (5.24) belongs to the following set:*

$$\mathcal{O} = \left\{ \left[\tilde{b}_i'\right]_{\tilde{b}_i}^{\min\{1,\tilde{b}_{i+1}\}} : i = 0, \cdots, L \right\}, \tag{5.27}$$

where $\tilde{b}_i'$ is defined as (5.26) and $[x]_a^b$ denotes $\min\{\max\{x, a\}, b\}$. In other words, $\theta^* = \arg\max_{\theta\in\mathcal{O}} d_{rate}^{(O)}$.

Proof. See Appendix 5.9.4. □

Parameters A_i, C_i, E_i and F_i can be calculated through partial sum, which leads to a computational complexity of $O(M)$ to get the set $\mathcal{O}$. Recalling that L is the number of D2D types with $\tilde{b}_i < 1$, the cardinality of $\mathcal{O}$, denoted by $|\mathcal{O}|$, is at most $L+1$, where $L \leq M$. Note that M is generally a small number, which implies that $L+1$ is small. Thus, Proposition 5.6 significantly reduces the complexity compared to the brute force search. Note that $E_i - F_i D$ decreases as θ increases. We have shown in Appendix 5.9.4 that the objective function is non-decreasing when $E_i - F_i D \geq 0$. Therefore, we only need to search over the domains where $E_i < F_i D$, and thus $|\mathcal{O}|$ can be further reduced.

5.6.2 Optimization of the Shared Network

In this section, we turn our attention to the optimization of the performance in the shared network. Similarly to the dedicated network, the objective is to maximize the utility function in terms of the rate lower bounds given by Corollary 5.3. We again consider a heavily loaded network with $\rho^{(S)} = 1$ and $p_a^{(S)} = \frac{7B\lambda_B}{9\left(b_C\lambda_U + \sum_{i=1}^{M} b_{D_i}(1-p_{t_i})\lambda_{D_i}\right)}$. Under these assumptions, we have the following conclusion.

Proposition 5.7. *Given $w \geq 1$, the optimal time hopping to maximize the rate density (5.25) is $p_{t_i}^* = 1,\ \forall i \in \{1, \cdots, M\}$. In contrast, when $w \to 0$, we have $p_{t_i}^* \to 0$.*

Proof. See Appendix 5.9.5. □

Similarly to the dedicated network, w can be adopted as a parameter to balance load between cellular and D2D networks, by decreasing which we can push D2D links to cellular mode. Though it is difficult to obtain the optimal frequency hopping in closed form in a general shared network, the maximization has been reduced to a lower-dimensional problem by finding the optimal time hopping probability, and the complexity to search the optimal scheme becomes much less. Denoting the number of possible values of p_{t_i} and p_{f_i} by $|p_t|$ and $|p_f|$, respectively, the complexity of brute force can be reduced from $\mathcal{O}((|p_t| \times |p_f|)^M)$ to $\mathcal{O}(|p_f|^M)$ (e.g., for the case with $|p_t| = |p_f| = 100$ and $M = 2$, the complexity is reduced from $\mathcal{O}(10^8)$ to $\mathcal{O}(10^4)$).

5.7 Performance Evaluation

In this section, we provide simulation results to validate the proposed model and analytical results. The main simulation parameters used in this chapter are summarized in Table 5.2, unless otherwise specified. The total bandwidth, noise power, path loss exponent, transmit power, and density of BSs are chosen based on 3GPP documents (see, e.g., [158, 159]). As for the other parameters, since the D2D traffic demand and its growth is not clear at this stage, the values are chosen given the best information available to us.

5.7.1 Validation of the System Model

We validate our analysis in Figs. 5.3 and 5.4. In Fig. 5.3, we compare the analytical SINR CDFs of D2D and cellular links in dedicated networks

Table 5.2: Simulation parameters for D2D-cellular networks with time-frequency hopping

Total bandwidth	10MHz
Number of sub-bands B	50
Number of D2D types M	2
Service demand of type i D2D links b_{D_i}	5, 15 subbands
Service demand of cellular users b_C	5 subbands
Density of BSs λ_B	$1/500^2$ m^{-2}
Density of cellular users λ_U	$60/500^2$ m^{-2}
Density of type i D2D links λ_{D_i} (same density for different types)	$15/500^2$ m^{-2}
Average distance between a D2D transmitter and receiver $\delta\sqrt{\frac{\pi}{2}}$	50 m
Transmit power of BSs P_B	46 dBm
Transmit power of D2D transmitters P_D	20 dBm
Noise power σ^2	-104 dBm
Path loss exponent α	3.5

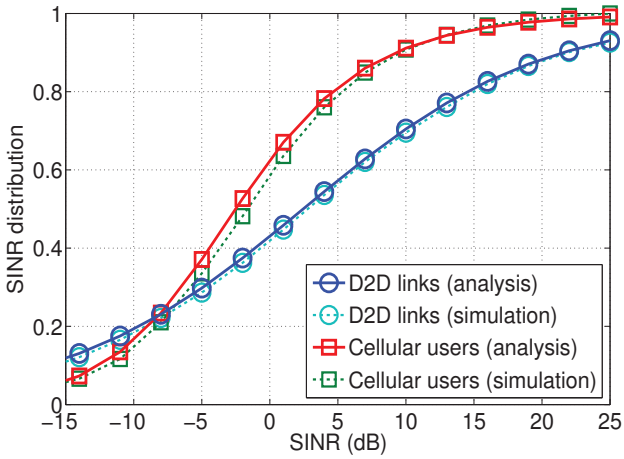


Figure 5.3: The SINR CDFs of active D2D links and cellular users in the dedicated network, with hopping probabilities $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.2$ and $p_{f_2} = 0.6$.

(given in Props. 5.1 and 5.2) to their corresponding simulations. The SINR CDFs of D2D and cellular links in shared networks (given in Theorems 5.2 and 5.3) are shown in Fig. 5.4. Recall that we approximate the set of interfering BSs by a PPP with density $\rho\lambda_B$. This approximation leads to gaps between the analysis and simulations (e.g., the gap between analytical and simulated SINR CDFs of cellular links in the dedicated network). However, the gaps are very small, which implies that the approximation is reasonable. From the fact that analytical results and their corresponding simulation results are in quite good agreement, we conclude that stochastic geometry allows us to efficiently find the approximate coverage probabilities for the D2D-enabled cellular network.

We validate the anaytical results of rates in Figs. 5.5 and 5.6. We fix the ratio of D2D density to cellular user density (e.g., 1/2), and increase these

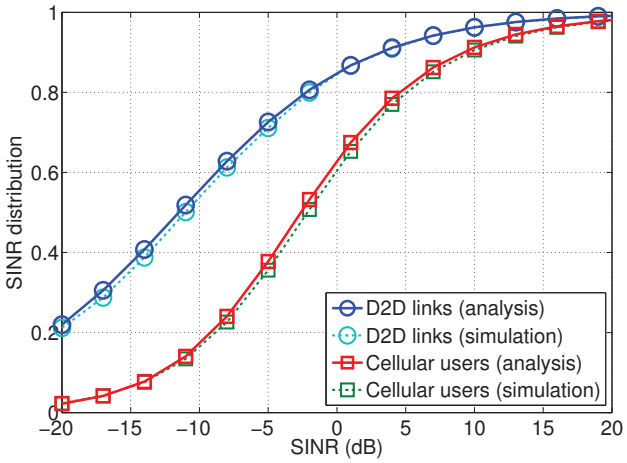


Figure 5.4: The SINR CDFs of D2D links and cellular users in the shared network. The hopping probabilities are $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.1$ and $p_{f_2} = 0.3$.

two densities proportionally. The analytical results are almost the same as the simulation results. The average rates of both cellular and potential D2D links decrease as the density increases, due to the decreasing available resources per link, as well as the increasing interference. Comparing the dedicated and shared networks, the D2D links have much higher average rate in the dedicated network. This implies that in a hybrid network sharing downlink resources, the interference from BSs may significantly degrade the network performance. We can observe that the D2D rates in the shared network first decrease very fast, and then much more slowly when the BSs become fully loaded. Indeed, when BSs are lightly loaded, the interference from BSs increases as the user density increases, which makes the D2D SINR decrease. In the fully loaded case, the interference from BSs stays almost the same. Though the interference from other D2D links increases, the decrease of D2D rate slows down, which implies

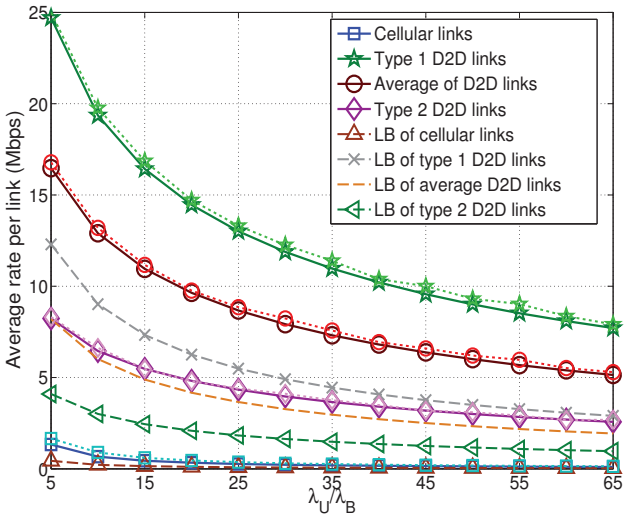


Figure 5.5: The average rates vs. the density of users in the dedicated network ($\theta = 0.5$). The hopping probabilities are $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.2$ and $p_{f_2} = 0.6$. The density of potential D2D links increases proportionally to the density of cellular users. The dashed lines are the simulation results while the solid lines are the corresponding analytical results.

that the interference from BSs is dominant in the performance of D2D links. Though the lower bound of rates are not very tight, the shapes are almost the same as the exact simulated rates, providing possibilities for optimization in terms of simple closed-form lower bounds. We compare the performance of exact rates and their lower bounds in the following subsection.

5.7.2 Optimization of Network Performance

The variation of rate density with time and frequency hopping probabilities in a heavily loaded network are shown in Figs. 5.7 and 5.8, respectively. As we can observe, the optimal hopping probabilities to maximize

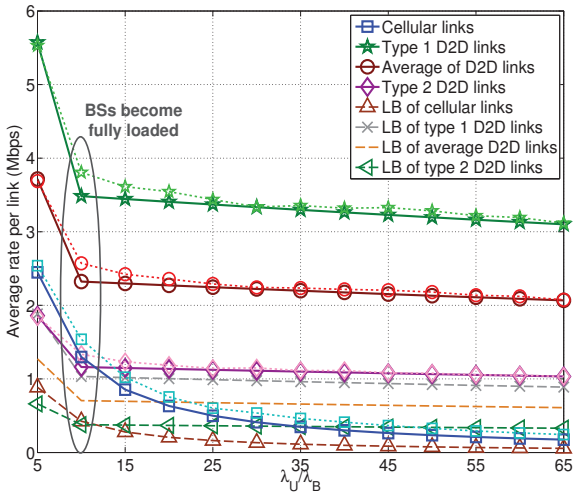


Figure 5.6: The average rates vs. the density of users in the shared network. The hopping probabilities are $p_{t_1} = p_{t_2} = 1$, $p_{f_1} = 0.1$ and $p_{f_2} = 0.3$. The dashed lines are the simulation results while the solid lines are the corresponding analytical results.

rate lower bounds are the same as the ones to maximize the exact rates. To maximize rate density, the active D2D links access the frequency resource according to their service demands in both dedicated and shared networks (i.e., $p^*_{f_i} = \min\{1, b_{D_i}/B_C\}$). All the potential D2D traffic is transmitted directly by D2D to alleviate the heavy load situation in the cellular network, and thus to maximize the total rate. Note that the optimal mode selection may be different for other objective functions. As shown in these two figures, the overall rate with dedicated allocation is greater than shared allocation in heavily loaded networks. One possible reason is that the rate of D2D links in the dedicated network is much larger than in the shared network, where the interference from BSs may limit the network performance, as it is observed

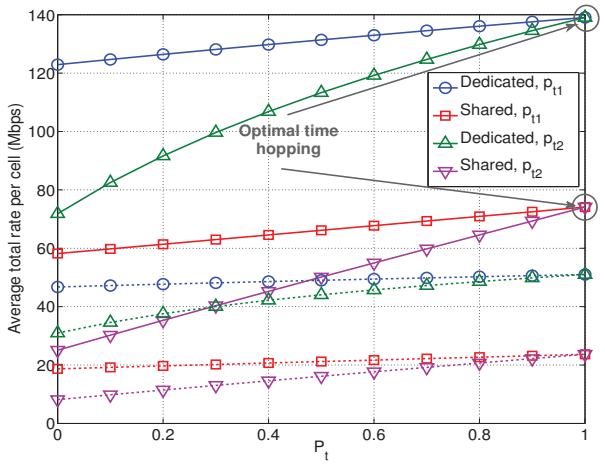


Figure 5.7: Effect of time hopping probabilities on the total rate density in heavily loaded networks ($\theta = 0.5$). The solid curves and dashed curves show the performance of exact rates and their lower bounds, respectively. We let frequency hopping probabilities be $p_{f_i} = \min\{1, B_C/b_{D_i}\}$.

in Figs. 5.5 and 5.6. By appropriately allocating resources between D2D and cellular networks, the active D2D links can get a quite large rate compared to cellular UEs. However, the shared network may overwhelm the dedicated network without optimal resource partition, which is investigated in Fig. 5.9.

Though in the setting of this chapter, the dedicated network has a greater rate than the shared network, the conclusion differs in different scenarios. For example, with an additional condition to guarantee the cellular network performance in the dedicated network (e.g., $\theta \leq 0.1$), the optimal total rate in the shared network would be greater than the total rate in the dedicated network, as illustrated in Fig. 5.9. Another example is the network with a small λ_D, which may have a better performance using the shared ap-

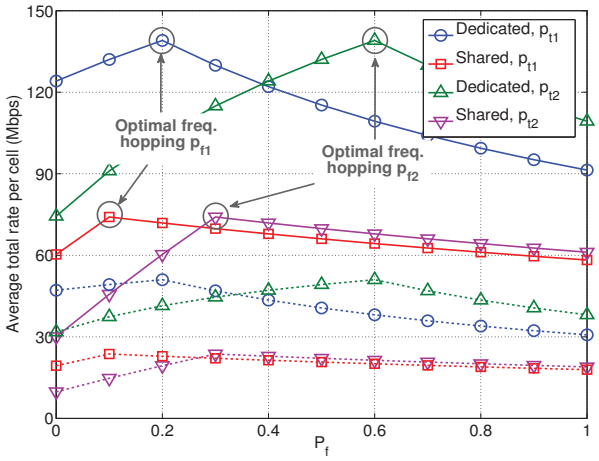


Figure 5.8: Effect of frequency hopping probabilities on the total rate density in heavily loaded networks ($\theta = 0.5$). The solid curves and dashed curves show the performance of exact rates and their lower bounds, respectively. We let the time hopping probabilities be $p_{t_i} = 1$.

proach (e.g., with $\lambda_D = 0.1\lambda_B$, the total rates per cell of dedicated and shared networks are 9.6 and 13.6 Mbps, receptively). Therefore, there is no absolute advantage either for the dedicated or shared approaches in general settings.

Fig. 5.9 also shows that the optimal resource partition in our simulation setup to maximize the total rate is $\theta^* = 1$. We can have different θ^* if the system parameters change. For example, Fig. 5.10 shows that $\theta^* = \frac{b_{D_2}}{B}$ when the average distance between the D2D transmitter and its receiver increases to 280m. This is consistent with the conclusion made in Proposition 5.6, where we claim that θ^* depends on various network parameters (e.g., δ) and belongs to the set $\mathcal{O}$. Note that we get the solution $\theta^* = 0$ or 1, which is unfair, due to that we consider the total rate maximization as our objective function for

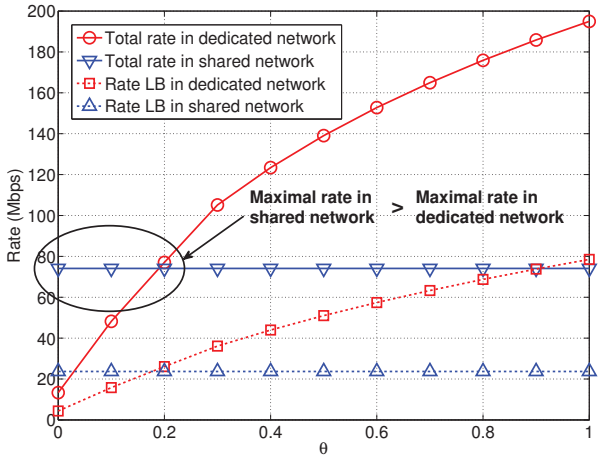


Figure 5.9: Total rate versus θ. The solid curves and dashed curves show the performance of exact rates and their lower bounds, respectively. We let frequency hopping probabilities be $p_{f_i} = \min\{1, B_C/b_{D_i}\}$ and the time hopping probabilities be $p_{t_i} = 1$.

the first-cut study. The θ^* would be very different if other utility functions are considered. For example, for maximization of log-rate, we will never get $\theta^* = 0$ or 1 (techniques similar to [28] can be used for this analysis). We leave the investigation of other utility functions to future work.

Though in most cases, we have $w \geq 1$, we investigate the impact of w on the optimal mode selection to maximize the total rate in Fig. 5.11. As w increases, which can be interpreted as the increasing price of cellular resources, the cellular communication becomes more and more unattractive for potential D2D traffic, and thus the load is shifted from cellular networks to D2D networks, in order to maximize the total rate. Therefore, it is possible to extend the current framework to a system, which can dynamically control

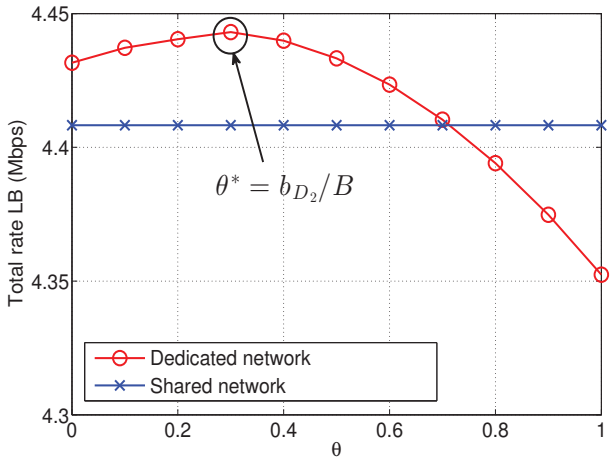


Figure 5.10: Rate versus θ in a network with the average distance between a D2D transmitter and its receiver being 280m. We let frequency hopping probabilities be $p_{f_i} = \min\{1, B_C/b_{D_i}\}$ and the time hopping probabilities be $p_{t_i} = 1$.

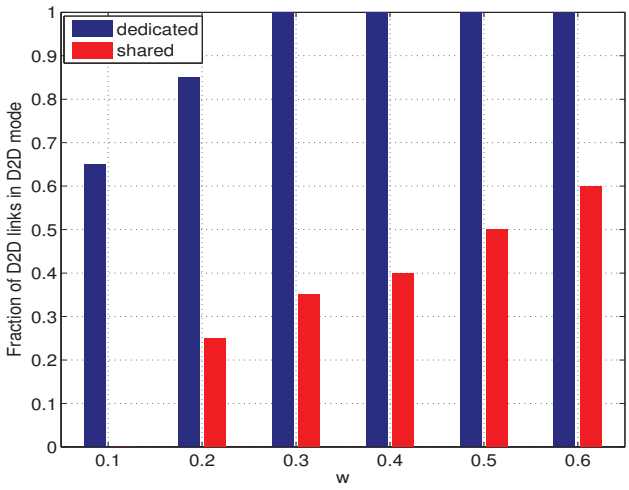


Figure 5.11: Effect of parameter w on the optimal mode selection to maximize the total rate.

w so as to adjust the load in D2D and cellular systems to achieve other more general utilities (e.g., utilities involving fairness). We leave the analysis to future work.

5.8 Summary

This chapter has presented tractable frameworks for both dedicated and shared networks, which provide accurate expressions for key performance metrics (e.g., coverage probability and average rate). With an appropriate resource partitioning, we observe that the dedicated network has a larger overall rate than the shared network in the downlink system. In dedicated networks, D2D links would access frequency bands as many as needed (i.e., $p^*_{f_i} = \min\{1, b_{D_i}/(\theta B)\}$) to maximize any non-decreasing utility function. To maximize the total rate, the potential D2D links are all in D2D mode in both fully loaded dedicated and shared networks, when $w \geq 1$. There are numerous extensions of the proposed flexible model, like multiple antennas, power control, interference management, more intelligent scheduling schemes and study of other utility functions. For example, one possible extension is to use the SIR-based CSMA protocol [143]. Though the set of active D2D links is no longer a homogeneous PPP, we can approximate it to a PPP with appropriate density, at little cost of accuracy. Then we can use the proposed model in this chapter to analyze the network performance. Another possible extension is to model BSs as other point processes, e.g., Matern hard core process (MHC), which characterizes the repulsiveness of BSs [160].

Recall that the optimized network throughput in this framework can serve as a lower bound for D2D-enable cellular networks with more sophisticated scheduling scheme. More dynamic resource allocations of D2D communication is discussed in the following chapter.

5.9 Appendix

5.9.1 Proof of Proposition 5.1

Conditioning on the distance between a typical transmitter and its receiver, we have

$$\begin{aligned}\mathbb{P}\left(\text{SINR} > \beta \mid v\right) &= \mathbb{P}\left(h_0 > s(I_{\tilde{\Phi}_D} + \sigma^2) \mid v\right) \\ &\overset{(a)}{=} \mathbb{E}_{I_{\tilde{\Phi}_D}}\left[\exp\left(-s(I_{\tilde{\Phi}_D} + \sigma^2)\right)\right] \\ &= e^{-s\sigma^2}\mathcal{L}_{I_{\tilde{\Phi}_D}}(s),\end{aligned}$$

where $s = \beta P_D^{-1} v^{\alpha}$, and $\mathcal{L}_{I_{\tilde{\Phi}_D}}(s)$ is the Laplace transform of random variable $I_{\tilde{\Phi}_D}$. The equality (a) follows from $h_0 \sim \exp(1)$, and the last equality follows from the independence of noise and interference. The Laplace transform can be further derived as follows:

$$\begin{aligned}\mathcal{L}_{I_{\tilde{\Phi}_D}}(s) &= \mathbb{E}\left[\exp\left(-s\sum_{Z_i \in \tilde{\Phi}_D \setminus 0} P_D h_i |Z_i|^{-\alpha}\right)\right] \\ &\overset{(a)}{=} \exp\left(-2\pi\tilde{\lambda}_D \int_0^{\infty}\int_0^{\infty}\left(1 - e^{-sP_D h/u^{\alpha}}\right) F(dh) u du\right) \\ &= \exp\left(-2\pi\tilde{\lambda}_D \mathbb{E}_h\left[\int_0^{\infty}\left(1 - e^{-sP_D h/u^{\alpha}}\right) u du\right]\right) \\ &\overset{(b)}{=} \exp\left(-\pi\tilde{\lambda}_D \mathbb{E}_h\left[\Gamma\left(1 - \frac{2}{\alpha}\right)(shP_D)^{\frac{2}{\alpha}}\right]\right) \\ &= \exp\left(-\tilde{\lambda}_D \frac{2\pi^2/\alpha}{\sin\left(2\pi/\alpha\right)}(sP_D)^{\frac{2}{\alpha}}\right),\end{aligned}$$

where $F(dh)$ is the law of channel fading (e.g., $F(dh) = e^{-h}dh$ in Rayleigh fading), and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. The equality (a) follows from the Slivnyak's Theorem of a PPP and the Laplace functional of a PPP [153, 154], (b) is obtained by changing $x = \frac{shP_D}{r^\alpha}$, and the last equality follows from the Rayleigh fading assumption. Then we complete the proof by deconditioning on v.

5.9.2 Proof of Corollary 5.2

In this special case, for D2D links, we have

$$\begin{aligned}\mathbb{P}_D(\beta) = &\int_0^\infty \exp\left(-\beta P_D^{-1}\sigma^2 v^\alpha - \tilde{\lambda_D}\frac{2\pi^2/\alpha\beta^{\frac{2}{\alpha}}v^2}{\sin(2\pi/\alpha)}\right.\\ &\left.-2\pi\lambda_B H_0(\beta,\alpha)v^2 - \frac{v^2}{2\delta^2}\right)\frac{v}{\delta^2}dv\\ =&\frac{1}{2\delta^2\tilde{\lambda_D}\frac{2\pi^2/\alpha}{\sin(2\pi/\alpha)}\beta^{\frac{2}{\alpha}} + 4\delta^2\pi\lambda_B H_0(\beta,\alpha)+1},\end{aligned}$$

where the last equality is obtained by letting $x = v^2$ and calculating the integral over x. As for the cellular users, according to (5.15), we have

$$\begin{aligned}\mathbb{P}_C(\beta) = &\int_0^\infty \exp\left(-\tilde{\lambda_D}\frac{2\pi^2/\alpha}{\sin(2\pi/\alpha)}(\beta\frac{P_D}{P_B})^{\frac{2}{\alpha}}r^2\right.\\ &\left.-2\pi\lambda_B H_1(\beta,\alpha)r^2 - \lambda_B\pi r^2\right)2\lambda_B\pi r dr\\ =&\frac{1}{\frac{\tilde{\lambda_D}}{\lambda_B}\frac{2\pi/\alpha}{\sin(2\pi/\alpha)}(\beta\frac{P_D}{P_B})^{\frac{2}{\alpha}} + 2H_1(\beta,\alpha)+1}.\end{aligned}$$

5.9.3 Proof of Proposition 5.5

Plugging $p^*_{f_i}$ to (5.24), the average rate of active D2D links in (5.24) is non-decreasing with respect to p_{t_i}. Denoting the average rate of cellular users

by $g(p_{t_i})$, we have

$$g = \frac{7(1-\theta)B\lambda_B}{9w} \frac{\log_2(1+\beta_C)}{2H_1(\beta_C,\alpha)+1} \frac{\left(\sum_{i=1}^{M} \lambda_{D_i} b_{D_j}(1-p_{t_j}) + w\lambda_U b_C\right)}{\left(\sum_{i=1}^{M} \lambda_{D_i} b_{D_i}(1-p_{t_i}) + \lambda_U b_C\right)},$$

whose first derivative with respect to p_{t_k} is

$$\frac{\partial g}{\partial p_{t_k}} = \frac{7(1-\theta)B\lambda_B}{9w} \frac{\log_2(1+\beta_C)}{2H_1(\beta_C,\alpha)+1} \frac{\lambda_{D_k} b_{D_k} \lambda_U b_C (w-1)}{\left(\sum_{i=1}^{M} \lambda_{D_i} b_{D_i}(1-p_{t_i}) + \lambda_U b_C\right)^2} \geq 0,$$

where the last inequality follows from the assumption that $w \geq 1$ for congested networks. Therefore, the rate density is a non-decreasing function of p_{t_i}, and thus $p_{t_i}^* = 1$.

As for a lightly loaded network, with w being very small, the second term dominates the rate density, which is non-increasing with respect to p_{t_i} when $w \to 0$. Therefore, $p_{t_i}^* = 0$.

5.9.4 Proof of Proposition 5.6

Denote $\beta_C = \arg\max_\beta R_{Cl}^{(O)}$ and $\beta_D = \arg\max_\beta R_{Dl}^{(O)}$. Plugging $p_{t_i}^* = 1$ to (5.24), the objective function becomes

$$\max_\theta \ \frac{\sum_i \lambda_{D_i} p_{f_i}^* \theta B \log_2(\beta_D+1)}{1 + 2\delta^2 \sum_i \lambda_{D_i} p_{f_i}^* \frac{2\pi^2/\alpha}{\sin(2\pi/\alpha)} \beta^{\frac{2}{\alpha}}} - \frac{7B\lambda_B \log_2(\beta_C+1)}{9\left(2H_1(\beta_C,\alpha)+1\right)}\theta. \tag{5.28}$$

Recall $A_i = \sum_{j\in\mathcal{S}_i} \lambda_{D_j} \tilde{b}_j B \log_2(\beta_D+1)$, $C_i = 2\delta^2 \sum_{j\in\mathcal{S}_i} \lambda_{D_j} \tilde{b}_j \frac{2\pi^2/\alpha}{\sin(2\pi/\alpha)} \beta^{\frac{2}{\alpha}}$, $D = \frac{7B\lambda_B}{9} \frac{\log_2(\beta_C+1)}{2H_1(\beta_C,\alpha)+1}$, $E_i = \sum_{j\in\mathcal{G}_i} \lambda_{D_j} B \log_2(\beta_D+1)$, and $F_i = 2\delta^2 \sum_{j\in\mathcal{G}_i} \frac{2\pi^2 \lambda_{D_j}/\alpha}{\sin(2\pi/\alpha)} \beta^{\frac{2}{\alpha}} + 1$. On the ith region, the objective function (5.28) can be written as

$$\max_\theta \ \frac{E_i\theta^2 + A_i\theta}{F_i\theta + C_i} - D\theta,$$

whose first derivative is $\frac{E_iF_i\theta^2+2C_iE_i\theta+A_iC_i}{(C_i+F_i\theta)^2} - D$ and the second derivative is $\frac{2C_i(E_iC_i-A_iF_i)}{(C_i+F_i\theta)^3}$. Note that $C_iE_i = A_i(F_i-1)$, i.e., $C_iE_i < A_iF_i$. We consider the following two cases.

(1) For partitions with $E_i \geq F_iD$, the first derivative of (5.28) is non-negative, and thus the objective function is non-decreasing. In this case, we have $\theta^* = \tilde{b}_{i+1}$. Note that $E_i - F_iD$ decreases as i increases. Denoting the index of the last domain that satisfies $E_i \geq F_iD$ by k, the objective function keeps increasing over the first k partitions, and thus $\theta^* = \tilde{b}_{k+1}$ for the first k partitions.

(2) For the partitions with $E_i < F_iD$, we have a positive second derivative, implying that the objective function is concave. Thus, the optimal solution in the latter case is $\left[\frac{1}{F_i}\left(\sqrt{\frac{C_i(A_iF_i-E_iC_i)}{DF_i-E_i}} - C_i\right)\right]_{\tilde{b}_i}^{\min\{1,\tilde{b}_{i+1}\}}$, where $[x]_a^b$ denotes $\min\{\max\{x,a\},b\}$.

Combining the above two cases, the proof is complete.

5.9.5 Proof of Proposition 5.7

When $p_{f_i}B > b_{D_i}$, the objective function is non-increasing with respect to p_{f_i}, and thus we have $p_{f_i}^* \leq b_{D_i}/B$. Observing that p_{f_i} only appears in terms of $p_{t_i}p_{f_i}$, we change variable to $x_i = p_{t_i}p_{f_i}$. The rate density maximization problem becomes

$$\begin{aligned} \max_{p_t,x} \quad & d_{\text{rate}}^{(S)}(x_i, p_{t_i}) \\ \text{s.t.} \quad & x_i \leq b_{D_i}/B, \\ & x_i \leq p_{t_i} \leq 1, \forall i \in \Phi_D. \end{aligned} \tag{5.29}$$

The objective function in (5.29) is

$$\begin{aligned} & d_{\text{rate}}^{(S)}(x_i, p_{t_i}) \\ =& \frac{\sum_{j=1}^{M} x_j \lambda_{D_j} B \log_2(1+\beta_D)}{\frac{2\pi^2/\alpha 2\delta^2 \beta_D^{\frac{2}{\alpha}}}{\sin(2\pi/\alpha)} \sum_{i=1}^{M} x_i \lambda_{D_i} + 4\delta^2 \pi \lambda_B H_0(\beta_D, \alpha) + 1} \\ & + p_a^{(S)} \frac{\left(\sum_{j=1}^{M} \frac{b_{D_j}}{w}(1-p_{t_j})\lambda_{D_j} + b_C \lambda_U\right) \log_2(1+\beta_C)}{\frac{\sum_{i=1}^{M} x_i \lambda_{D_i}}{\lambda_B} \frac{2\pi/\alpha}{\sin(2\pi/\alpha)} (\beta_C \frac{P_D}{P_B})^{\frac{2}{\alpha}} + 2H_1(\beta_C, \alpha) + 1}, \end{aligned}$$

where the first term is independent of p_{t_i}, and the second term can be written as

$$A \frac{\left(\sum_{j=1}^{M} b_{D_j}(1-p_{t_j})\lambda_{D_j} + w b_C \lambda_U\right)}{\left(b_C \lambda_U + \sum_{i=1}^{M} b_{D_i} \lambda_{D_i}(1-p_{t_i})\right)}, \tag{5.30}$$

where $A = \frac{\frac{7B\lambda_B}{9w}}{\frac{\sum_{i=1}^{M} x_i \lambda_{D_i}}{\lambda_B} \frac{2\pi/\alpha}{\sin(2\pi/\alpha)} (\beta_C \frac{P_D}{P_B})^{\frac{2}{\alpha}} + 2H_1(\beta_C, \alpha) + 1} > 0$. The first derivative of (5.30) with respect to p_{t_i} is

$$A \frac{b_C \lambda_U b_{D_i} \lambda_{D_i} (w-1)}{\left(b_C \lambda_U + \sum_{i=1}^{M} b_{D_i} \lambda_{D_i}(1-p_{t_i})\right)^2},$$

which is non-negative when $w \geq 1$. Therefore, given $w \geq 1$, the objective function (5.25) is a non-decreasing function of p_{t_i}, and we have $p_{t_i}^* = 1$.

In a lightly loaded network with small w, similar to the proof in Appendix 5.9.3, we have $p_{t_i}^* = 0$.

Chapter 6

Distributed Resource Allocation in D2D Enhanced Cellular Networks[1]

As established in Chapter 5, the strong interference from BSs to D2D communication may kill the throughput gain when D2D links reuse the downlink cellular resources. On the other hand, [28] as a parallel work to Chapter 5 shows that the dedicated and shared resource allocation have comparable throughput in the context of uplink cellular systems. Compared to downlink, the uplink resources are often under-utilized [29]. Moreover, when D2D links reuse uplink resources, the interference from D2D to cellular transmissions can be better handled, since the BSs that are more powerful than UEs suffer from D2D interference. Therefore, sharing uplink spectrum is preferable overall [3]. This motivates the study of D2D communication in the context of cellular uplink systems in this chapter, where the resources are shared between D2D and cellular links to improve the resource utilization. In such cases, the success of co-existence of D2D and cellular transmissions depends heavily on the interference management.

[1]This chapter has been published in [161]. Coauthor Dr. Mazin Al-Shalash has provided technical suggestions and many insights to this work, and Dr. Constantine Caramanis and Dr. Jeffrey G. Andrews are my supervisors.

The objective of this chapter is to improve the network throughput by allowing D2D communication to share the cellular uplink resources. At the same time, we restrict the access of D2D links to the uplink spectrum in order to manage the interference. Generally, centralized interference management requires a central controller (e.g., the BS) to acquire the CSI between each transmitter and receiver. This requires high overhead, particularly in the scenario where channels vary rapidly (e.g., in a high-mobility environment). Therefore, a distributed algorithm requiring only local information is preferable. We address the following design question in this chapter: how to intelligently manage spectrum for D2D with only local information and BS assistance (e.g., setting a high cost for D2D links that cause strong interference), so as to manage the interference and improve the network throughput?

6.1 Related Work

There has been increasing interest recently in the investigation of interference management in shared networks. Power control is one viable approach, e.g., consideration of a greatly simplified model with one cellular UE and one D2D link [126], a simple power reduction method based on the derived SINR [162], and study of several power control schemes including fixed power and fixed SNR target [163] are some existing works. Another popular approach, related to the direction we propose, is to intelligently manage spectrum for D2D links based on channel conditions and nearby interfering UEs, e.g., maximal mutual interference minimization [164], network through-

put maximization [129], setting exclusive D2D transmission zone to achieve interference avoidance [127, 128], and interference randomization through frequency and/or time hopping in [134] and Chapter 5 [123]. However, the key new aspect of D2D-enabled cellular networks – BS assistance – has received much less treatment in the literature. Possibly one reason for this is that the computational problem itself is quite difficult, and the communication and coordination alone required for a good centralized solution might be prohibitive.

As we discuss below, the key approach adopted in this chapter is a two-stage distributed algorithm that has a game theoretic interpretation: BSs send out a signal representing a fictional price that can be considered as the assistance from BSs, and then D2D users optimize a local objective function adapting to that price and to what the other users are doing. Since individual D2D links optimize their local functions "selfishly", this approach has a game theoretic interpretation, which allows us to use algorithms and concepts from game theory, even though there is no actual market, and users agree to "play" this game using the BS's signal, without actually exchanging currency.

Game theory has been used in various disciplines to model competition for limited resources in more general networks. For example, work has been done considering spectrum sharing based on local bargaining [165], repeated game [166], auction mechanisms [132] and two-stage game [133, 167]. Paper [168] demonstrates several different game models for D2D resource allocation, where an interesting example is to use the reverse iterative combinatorial auction [169]. In this chapter, the game theoretic approach is used as

an *algorithmic technique* to obtain efficient distributed spectrum management. Similar to the recent work [133], we model a Stackelberg game to control the interference from D2D to cellular network. The key difference from [133] is in the upper-stage problem, where we take into account the D2D rate. Moreover, we investigate the convergence of the algorithms for both lower-stage and upper-stage problems. Works in cognitive radio such as [167] are related to the second part of our work (i.e., the investigation of optimal prices to charge D2D links accessing the shared resources), which proposes that secondary users adapt their powers for alleviating interference to primary users. The key techniques used in this chapter to study the optimal prices are similar to [167], but we in addition investigate the convergence of the spectrum management scheme for the D2D network (i.e., the lower-stage problem), as well as the convergence of the proposed heuristic algorithm.

6.2 Contributions

In this chapter, we present a distributed, efficient and low-overhead spectrum management method for D2D links to improve the throughput while keeping the performance of cellular users at a guaranteed level. Specifically, the main contributions are:

A Two-stage distributed algorithm. We propose an iterative two-stage algorithm in Section 6.4. In the first stage, the BSs send a pricing signal that adapts to the gap between the aggregate interference from D2D links and a predefined interference tolerance level, where the price increases if D2D

interference is higher than the tolerance level and decreases otherwise. In the second stage, each D2D link independently maximizes its utility consisting of a reward equal to its expected rate and a penalty proportional to the interference caused by this link to the BS, as measured by the pricing signal. Note that this two-stage model is a Stackelberg game [170], and the algorithm can be seen as a pricing mechanism. This algorithm requires no cooperation among D2D links, yet succeeds in discouraging strongly interfering or low-SINR D2D links to access more RBs.

Utility-based D2D resource allocation adaptation. In Section 6.5, we consider the lower-stage problem, where we maximize the D2D rate in terms of expected SINR for tractability, which provides a performance upper bound and can serve as a benchmark. Each D2D link selfishly maximizes its utility given other D2D links' decisions and the price broadcast from BSs, which essentially forms a non-cooperative game. To reduce the computational complexity and overhead, we further consider the problem of maximizing a lower bound of the utility function for each D2D link. We then propose an efficient iterative algorithm similar to a waterfilling algorithm, which only requires local information. Our simulation results show that the result obtained by the proposed iterative algorithm is very close to the solution to the upper bound problem. This further lightens the computational burden on each user.

Cellular link performance protection. Given the solution of the lower-stage problem, we study the optimal price the BSs report in Section 6.6, to maximize the network utility while protecting cellular links. We show

that this problem can be transformed into a linear complementarity problem (LCP). This allows us to take advantage of, and adapt for our problem, general algorithms for LCP. We further propose a simpler heuristic algorithm based on the bisection method, and observe that it has low overhead and converges very quickly with almost no loss. We also propose a simple greedy algorithm that leads to efficient computation at the cost of overall throughput, where the throughput loss decreases as the interference tolerance level increases, e.g., the throughput loss compared to the algorithm for LCP are about 25% and 5% when the interference tolerance level is 5dB lower and above the cellular signal in our setup, respectively.

Numerical results in Section 6.7 show that the cellular links can be well protected with the average D2D throughput reduction of only 12% in our setup, compared to the scenario where all D2D links are active. On the other hand, compared to conventional cellular networks without D2D communication, the proposed algorithms provide significant throughput gain (about 5x with 10 D2D links per cell and average D2D link length 80m in our simulation setup). Note that the throughput gain highly depends on the amount of D2D traffic and average D2D link length. We take 10 D2D links per cell and link length 80m as an illustration example in this chapter.

6.3 System Model

We consider a uplink shared network, where cellular UEs in the same cell get different subbands (i.e., orthogonal chunks of RBs). Any general

scheduling scheme for cellular UEs can be used. Recall that a potential D2D link can either transmit directly by D2D communication, or transmit to a BS (i.e., mode selection). By intelligently conducting mode selection, we can adjust the aggregate interference in the network and thus optimize the achievable network performance. However, the mode selection variables in the SINR expression result in non-convexity of objective functions that are in terms of rate, i.e. $\log(1+\text{SINR})$. Moreover, the mode selection variables are binary, making the problem combinatorial. Note that different mode selection schemes lead to different optimal spectrum management, due to the differences in the resource allocation of cellular users and thus the differences in the interference tolerance level. On the other hand, spectrum management affects the achievable rate and thus affects the mode selection of D2D links. As discussed above, it is difficult to find the optimal mode selection, let alone the joint optimal mode selection and spectrum management of D2D links, where mode selection and spectrum management are coupled with each other. Despite the intractability of the optimization problem, there are various practical (but not necessarily optimal) mode selection approaches (e.g., distance-based mode selection [28]).

We propose the following mode selection as one viable scheme. The potential D2D transmitters are treated the same as cellular UEs when scheduling, except that we can add a weight to the scheduling metric. For example, with proportional fair scheduling [50], the user with the largest $qR_i/\bar{R}_i$ would be scheduled, where q is the weight, R_i and $\bar{R}_i$ are the instantaneous rate and average rate of link i, respectively. Without loss of generality, we assume cel-

lular users have $q = 1$. A typical value for q of potential D2D links might be $1/2$, since a potential D2D link in cellular mode would occupy both uplink and downlink resources. We can also let the weight q impose the cost on the backhual usage of core networks. Considering that D2D mode has more efficient resource utilization, the potential D2D links are biased against cellular mode using $q < 1$. Note that other mode selection schemes can also be applied to our following framework easily.

We assume potential D2D links that are not scheduled by BSs would be in D2D mode. In other words, we propose to let each BS complete the mode selection of potential D2D links in its coverage area. Given the mode selection, we aim to find the optimal spectrum management of D2D links to maximize the network utility. We leave the joint optimization of mode selection and spectrum management to future work.

Assuming that the cellular resource allocation is done by the BSs, our focus is on the spectrum management of D2D links. In this chapter, we consider resource allocation at each RB to simplify the notation and explanation, but any general units of RBs can be considered similarly. The sets of cellular UEs accessing the kth RB and of D2D links are denoted by $\mathcal{C}_k$ and $\mathcal{D}$, respectively, where the set of cellular UEs includes the potential D2D links in cellular mode. We define $\mathcal{I}_k$ as the set of D2D links accessing RB k (i.e., the set of interfering D2D links). Then the SINRs of D2D link i at RB k and a

cellular UE belonging to $\mathcal{C}_k$ are, respectively,

$$\text{SINR}_{ik}^{(D)} = \frac{\mathbb{1}_{\{i\in\mathcal{I}_k\}} P_{D_i} h_{ii}^{(k)}}{\sum_{j\in\mathcal{I}_k, j\neq i} P_{D_j} h_{ji}^{(k)} + \sum_{j\in\mathcal{C}_k} P_{C_j} h_{ji}^{(k)} + W_{ik}},$$

$$\text{SINR}_{ik}^{(C)} = \frac{P_{C_i} g_{ii}^{(k)}}{\sum_{j\in\mathcal{I}_k} P_{D_j} g_{ji}^{(k)} + \sum_{j\in\mathcal{C}_k, j\neq i} P_{C_j} g_{ji}^{(k)} + W_{ik}},$$

where $\mathbb{1}_{\{a\in\mathcal{A}\}}$ is an indicator function whose value is 1 if $a \in \mathcal{A}$ and 0 otherwise, P_{D_i} and P_{C_i} are the transmit powers of D2D and cellular links, respectively, $h_{ji}^{(k)}$ is the channel gain from UE j to D2D receiver i at RB k, $g_{ji}^{(k)}$ is the channel gain from UE j to the BS serving cellular UE i, and W_{ik} is the noise power of link i at RB k. We use Shannon capacity to calculate rate, i.e., $R_{ik} = B\log_2(1 + \text{SINR}_{ik})$, where B is the frequency bandwidth of a RB.

6.4 Problem Formulation

In this section, we first formulate a single-stage optimization problem to maximize the D2D throughput with a performance protection for cellular links. The computational intractability of the single-stage optimization then motivates us to consider a distributed setting, where each D2D link tries to maximize its own utility based only on local information.

6.4.1 Single-stage Problem Formulation

Without loss of generality, we let x_{ik} be the probability that D2D link i accessing RB k. The investigation of optimal access probabilities upper bounds the channel assignment problem where D2D links either access a RB or not

(i.e., the access probability is either 1 or 0). We consider the following utility maximization problem, subject to a D2D interference constraint to guarantee the cellular performance:

$$\begin{aligned} \max_{\mathbf{x}} \ & \sum_{i\in\mathcal{D}} w_i \sum_{k=1}^{K} R_{ik}^{(D)}(\mathbf{x}) \\ \text{s.t.} \ & \sum_{i\in\mathcal{D}} x_{ik} P_{D_i} g_{ii}^{(k)} \leq Q_k, \forall k, \\ & x_{ik} \in [0,1], \end{aligned} \tag{6.1}$$

where w_i is the weight for the ith D2D link, and K is the number of total available RBs for D2D links. Denoting the power set of $\mathcal{D}$ by $2^{\mathcal{D}}$, the rate of D2D link i at RB k is

$$R_{ik}^{(D)} = \sum_{\mathcal{I}_k\in 2^{\mathcal{D}}} \prod_{j\in\mathcal{I}_k} x_{jk} \prod_{n\in\mathcal{D}\setminus\mathcal{I}_k} (1-x_{nk}) \log_2\left(1+\mathrm{SINR}_{ik}^{(D)}\right). \tag{6.2}$$

The first constraint in (6.1) is for protecting cellular transmissions, where Q_k – called the *interference tolerance level* – depends on the channel condition of cellular links on RB k (e.g., Q_k could be the signal strength of the cellular link using RB k multiplied by a predefined threshold). Note that Q_k can be optimized to maximize a utility function incorporating the cellular rate. In this chapter, we consider Q_k as a predefined parameter and leave the joint optimization of Q_k and D2D resource allocation to future work. We observe that (6.1) is not a convex optimization problem. The computational complexity of a brute-force approach to solve (6.1) is $O(N_x^{N_D} N_D^2)$, where N_x is the number of possible values of x to be searched and N_D is the number of D2D links. Thus, the computation is essentially impossible for even a modest-sized network.

Instead of the centralized approach, we adopt a different strategy that results in an *efficient, distributed algorithm with low coordination, cooperation and communication overhead*. For tractability, we introduce variables – called *prices* for accessing RBs – to decouple the interference constraint in (6.1) and develop a distributed tractable framework. Particularly, BSs adjust prices to control the total D2D interference, and each D2D link individually maximizes its utility in terms of the expected rate and prices charged by BSs. This leads to a two-stage optimization problem, which consists of a problem to find optimal prices and several small-size convex optimization problems for D2D links. Though solutions to the two-stage problem may not provide the optimal solution to the original single-stage problem (6.1), this relaxation allows us to efficiently allocate resources in a distributed fashion, and the numerical results in Section 6.7 demonstrate a large rate gain without serious degradation in cellular performance using the proposed algorithm for the two-stage problem.

6.4.2 Two-stage Problem Formulation

We propose a pricing mechanism, where a BS charges the D2D link i in its coverage area the amount μ_{ik} per unit of the interference caused by this D2D link to the BS at RB k, i.e., the cost for a D2D link to access RB k is $\mu_{ik} x_{ik} P_{D_i} g_{ii}^{(k)}$. Assuming that each cell runs this mechanism independently, the cost of a D2D link only depends on the interference caused by this D2D link to its associated BS.

We assume the interference from other cells is invariant when we con-

sider the resource allocation in a typical cell. Therefore, we can incorporate the interference from neighboring cells into noise and the multi-cell scenario is simplified to a single-cell scenario. Under this assumption, the interference constraint is for the interference caused by D2D links in this cell. Note that in this case, the updated noise (incorporating inter-cell interference) is different from user to user, where generally cell-edge users suffer larger noise. Though we focus on the asynchronous scheduling scenario, the proposed framework can be easily generalized to a synchronous multi-cell scenario if the price at each RB is unified among different cells, where the BS in the proposed model becomes a network controller, and the interference becomes the aggregate interference from D2D links to all BSs in the network.

The net utility of D2D link i is $U_i = w_i \sum_{k=1}^{K} \left(R_{ik}^{(D)}(\mathbf{x}) - \mu_{ik} x_{ik} P_{D_i} g_{ii}^{(k)} \right)$, where the first and second term can be considered as the reward and penalty functions, respectively. The problem involves a non-cooperative network, where each D2D link aims to maximize its utility selfishly. We denote the access probabilities of D2D link i by $\mathbf{x}_i := [x_{i1}, x_{i2}, \cdots, x_{iK}]^T$. The access probabilities of all other D2D links are denoted by

$$\mathbf{x}_{-i} := [\mathbf{x}_1^T, \cdots, \mathbf{x}_{i-1}^T, \mathbf{x}_{i+1}^T, \cdots, \mathbf{x}_{N_D}^T]^T,$$

where N_D is the number of D2D links. Similarly, we define the *price vector* of D2D link i as $\boldsymbol{\mu}_i := [\mu_{i1}, \mu_{i2}, \cdots, \mu_{iK}]^T$. Given $\boldsymbol{\mu}_i$ and $\mathbf{x}_{-i}$, the problem for the D2D link i is

$$\begin{aligned} &\max_{\mathbf{x}_i} \; U_i(\mathbf{x}_i; \mathbf{x}_{-i}, \boldsymbol{\mu}_i) \\ &\text{s.t. } \; x_{ik} \in [0, 1], \; \forall k. \end{aligned} \tag{6.3}$$

On the other hand, the network aims to find optimal prices:

$$\begin{aligned} &\max_{\boldsymbol{\mu}\geq 0} U_c\left(\boldsymbol{\mu}, \mathbf{x}^*\left(\boldsymbol{\mu}\right)\right) \\ &\text{s.t. } \sum_{i\in\mathcal{D}} x_{ik}^*(\boldsymbol{\mu}) P_{D_i} g_{ii}^{(k)} \leq Q_k, \ \forall k, \end{aligned} \tag{6.4}$$

where $U_c\left(\boldsymbol{\mu}, \mathbf{x}^*\left(\boldsymbol{\mu}\right)\right) = \sum_{i\in\mathcal{D}} \sum_{k=1}^{K} \mu_{ik} x_{ik}^*\left(\boldsymbol{\mu}\right) P_{D_i} g_{ii}^{(k)}$, and $\mathbf{x}^*\left(\boldsymbol{\mu}\right)$ is the solution of (6.3) for a given $\boldsymbol{\mu}$. Taking a game theoretic perspective, the above problem is a decentralized Stackelberg game (a two-stage game), where the leader moves first and then the followers move accordingly. In this chapter, the BS is the leader and the D2D links are the followers.

To solve the two-stage problem, we use a backward induction technique. We start with the problem of the D2D links – called a *lower problem* – and get the D2D access probability $\mathbf{x}^*\left(\boldsymbol{\mu}\right)$. By plugging $\mathbf{x}^*\left(\boldsymbol{\mu}\right)$ into (6.4), we then investigate the network utility maximization – called an *upper problem*.

6.5 Lower Problem: A Non-cooperative D2D Network

Given $\boldsymbol{\mu}$, D2D links try to maximize their utility selfishly. This defines a non-cooperative game $G_D = [\mathcal{D}, \{\mathbf{x}_i\}, \{U_i\}]$.

For tractability, we use Jensen's inequality and consider the following objective function that upper bounds (6.2):

$$\begin{aligned} &\max_{\mathbf{x}_i} w_i \sum_{k=1}^{K} \tilde{R}_{ik}^{(D)}(\mathbf{x}) - \sum_{k=1}^{K} \mu_{ik} x_{ik} P_{D_i} g_{ii}^{(k)} \\ &\text{s.t. } x_{ik} \in [0,1], \ \forall k, \end{aligned} \tag{6.5}$$

where

$$\tilde{R}_{ik}^{(D)} = \log\left(1 + \sum_{\mathcal{J}_k \in 2^{\mathcal{D}}} \prod_{j \in \mathcal{J}_k} x_{jk} \prod_{n \in \mathcal{D} \setminus \mathcal{J}_k} (1 - x_{nk}) \mathrm{SINR}_{ik}^{(D)}\right).$$

The upper bound is tight if most x_{ik} are binary. We compare the gap between the solution maximizing (6.2) and (6.5) numerically in Section 6.7 and leave the analysis to future work.

We adopt an identical price for D2D links accessing the same RB, i.e., $\mu_{ik} = \mu_{jk}$. The rationale for doing this is that the BS only cares about the aggregate interference, rather than the differences between the interference values from different D2D links. The structure of (6.5) suggests decoupling the lower problem into K subproblems, where we consider each RB independently. In the rest of this chapter, we consider a typical RB, and ignore the RB index k for notation simplicity.

6.5.1 Distributed Algorithm Design

Optimization problems produce solutions with certain optimality guarantees. In our setting, however, the D2D links behave in a non-cooperative fashion. Thus, understanding the behavior and performance of our algorithm requires consideration of a different solution concept. This notion has been well-studied in game theory, and it is known that the analog of stationary points in an optimization solution are the so-called *Nash Equilibrium* (NE) points. In our context, these are the fixed points from which no D2D link would want to *unilaterally deviate* [170]. In the rest of this chapter, the NE points always refer to the NE of the D2D non-cooperative game G_D. In this

subsection, we study what these NE points are and propose an algorithm that converges to a NE.

We denote the feasible region of x_i by $\mathcal{X}_i$, where $\mathcal{X}_i = \{x_i \in [0,1]\}$. The existence of NE for the non-cooperative game is given by Lemma 6.1, according to the Debreu-Glicksberg-Fan Theorem [171–173].

Lemma 6.1. *If $\mathcal{X}_i$ is compact and convex, U_i is concave in x_i given $\mathbf{x}_{-i}$ and continuous, then the NE exists.*

It is straightforward to show that the above conditions are satisfied, and thus we have at least one NE. Then a natural question follows: how to attain a NE?

For fixed $\mathbf{x}_{-i}$ and μ, the problem (6.5) is a convex optimization, and the optimal solution is the point which vanishes the first derivative of the objective function (if feasible):

$$x_i^* = \left[\frac{w_i}{\mu P_{D_i} g_{ii} \ln 2} - \frac{1}{\sum_{\mathcal{I} \in 2^{\mathcal{D}}} \prod_{j \in \mathcal{I} \setminus i} x_j \prod_{n \in \mathcal{D} \setminus \mathcal{I}} (1 - x_n) \mathrm{SINR}_i^{(D)}} \right]_0^1, \tag{6.6}$$

where $[x]_0^1 = \min\{1, \max\{0, x\}\}$. We define the following function:

$$f(x_1, \cdots, x_{N_D}; \mu) = \left(x_1^*(\mathbf{x}_{-1}), \cdots, x_{N_D}^*(\mathbf{x}_{-N_D}) \right),$$

where $x_i^*(\mathbf{x}_{-i})$ is given by (6.6). Function f describes the optimal resource access probabilities given that the access probabilities of other links are fixed, and thus is called the *best-response (BR) function*. We propose a synchronous

iterative algorithm – called the *BR Algorithm*, where all D2D links adjust their access probabilities simultaneously according to

$$(x_1(t+1), \cdots, x_{N_D}(t+1)) = f(x_1(t), \cdots, x_{N_D}(t); \mu).$$

Applying the *Maximum Theorem* [174], we can show that f is continuous. Note that the BR Algorithm will never converge to a solution that is not a NE, since each D2D link has the access probability that maximizes its utility, which implies that no links can gain by changing only their own access probabilities unilaterally at the convergence point.

Though procedures of the BR Algorithm are simple, the complexity to calculate (6.6) is high, due to the expectation calculation involving N_D Bernoulli random variables, whose complexity is $O(2^{N_D} N_D^2)$. In addition, D2D links need to exchange their current access probabilities, which causes high overhead. The overhead and complex computation are not desirable, especially for UEs that are power limited. Other algorithms such as gradient-projection based algorithm [54] or algorithms in learning automata [175] can also be applied, with the disadvantages of either slow convergence or memory space limit. These motivate the following subsection, where we consider a lower bound of the objective function in (6.5).

6.5.2 Joint Resource Allocation and Power Control – A Lower Bound Problem

In problem (6.5), each D2D link maximizes the utility in terms of the expected SINR. Approximating the rate to be calculated by expected interfer-

ence rather than expected SINR, we have

$$\begin{aligned} \max_{x_i} \; & w_i \log\left(1+\mathrm{SINR}_i'\right) - \mu x_i P_{D_i} g_{ii} \\ \text{s.t.} \; & 0 \le x_i \le 1, \end{aligned} \tag{6.7}$$

where $\mathrm{SINR}_i' = \frac{x_i P_{D_i} h_{ii}}{\sum_{j\in\mathcal{D}, j\neq i} x_j P_{D_j} h_{ji} + \sum_{j\in\mathcal{C}} P_{C_j} h_{ji} + W_i}$. This problem motivates a low-complexity low-overhead algorithm, as shown below.

Variable x_i in (6.7) can be considered as a joint resource allocation and power control variable, where $\mathbb{1}\left(x_i > 0\right)$ indicates whether D2D link i accesses the RB, and the value of x_i denotes the fraction of maximal transmit power to use. The strategy with respect to (6.5) can be considered as a scheme similar to random hopping (with different hopping probabilities at each link), while the strategy in (6.7) is deterministic, which considers power control in addition to resource allocation. Intuitively, the hopping scheme randomizes strong interference, and thus may potentially provide a larger gain than the latter case, though we consider power control jointly. We show this relationship mathematically in Proposition 6.1.

Proposition 6.1. *The optimization problem (6.7) maximizes a lower bound of the utility function in (6.5).*

Proof. Denoting the interference from other D2D links by I, the SINR can be written as $\mathbb{E}\left[\mathrm{SINR}_{D_i}\right] = \mathbb{E}_I\left[\frac{P_{D_i} h_{ii}}{I + \sum_{j\in\mathcal{C}} P_{C_j} h_{ji} + W_i}\right]$. It is straightforward to verify that $f(I) = \frac{P_{D_i} h_{ii}}{I + \sum_{j\in\mathcal{C}} P_{C_j} h_{ji} + W_i}$ is convex. By Jensen's inequality, we have $f(\mathbb{E}[I]) \le \mathbb{E}\left[f(I)\right]$, which completes the proof. □

We call (6.7) the lower bound problem of (6.5) in this chapter. Invoking Lemma 6.1 again, we can show that there is at least one NE for the D2D game formulated in this subsection. Though there may exist multiple NEs in general, our setup admits a unique NE under some specific conditions; we specify those precisely in Section 6.5.3. Note that the NEs of the games with (6.5) and with (6.7) are not necessarily the same, and thus Proposition 6.1 does not say that the BR Algorithm in Section 6.5.1 always performs better than the algorithms proposed in the following subsection.

Given $\mathbf{x}_{-i}$ and μ, (6.7) is a convex optimization problem and its optimal solution is given by Proposition 6.2.

Proposition 6.2. *The solution of (6.7) has the following form*

$$x_i^* = \left[\frac{a_i - s_i}{P_{D_i} h_{ii}}\right]_0^1, \tag{6.8}$$

where $a_i = \frac{w_i P_{D_i} h_{ii}}{\mu P_{D_i} g_{ii}} - \sum_{j\in\mathcal{C}} P_{C_j} h_{ji} - W_i$, $s_i = \sum_{j\neq i} x_j P_{D_j} h_{ji}$, *and* $[x]_0^1 = \min\{1, \max\{x, 0\}\}$.

Proof. According to the KKT conditions [53], we have $\frac{\partial U_i}{\partial x_i} = 0$ if $x_i \in (0,1)$, $\frac{\partial U_i}{\partial x_i} \leq 0$ if $x_i = 0$, and $\frac{\partial U_i}{\partial x_i} \geq 0$ otherwise, where

$$\frac{\partial U_i}{\partial x_i} = \frac{w_i P_{D_i} h_{ii}}{x_i P_{D_i} h_{ii} + \sum_{j\in\mathcal{D}, j\neq i} x_j P_{D_j} h_{ji} + \sum_{j\in\mathcal{C}} P_{C_j} h_{ji} + W_i} - \mu_i P_{D_i} g_{ii}.$$

The above equations and inequations result in (6.8). □

Eq. (6.8) is similar to the waterfilling function in power allocation problems, except that our constraint $x_i \in [0,1]$ is independent over different

RBs and thus we obtain a closed-form solution (6.8). Leveraging existing works on waterfilling problems, we propose an iterative algorithm similar to the iterative waterfilling algorithm (see, e.g., [174, 176, 177]).

6.5.3 Algorithm Design for the Lower bound Problem

Similar to Section 6.5, we propose a synchronous iterative algorithm based on the BR function, defined as

$$f_L(x_1, \cdots, x_{N_D}; \mu) = \left(x_1^*(\mathbf{x}_{-1}), \cdots, x_{N_D}^*(\mathbf{x}_{-N_D})\right),$$

where $x_i^*(\mathbf{x}_{-i})$ is given by (6.8). The algorithm – called the *LB Algorithm* – is given by Algorithm 1. Similar to the BR Algorithm, we have that if the LB Algorithm converges, then it converges to a NE.

Algorithm 1 LB Algorithm: an iterative algorithm for lower bound problem of D2D

1: Initialization: given price $\mu \geq 0$, let $x_i(0) = 1, \forall i$, and $t = 0$;
2: **while** $\|\mathbf{x}(t) - \mathbf{x}(t-1)\| \geq \epsilon$ **do**
3: let $\mathbf{x}(t+1) = f_L(\mathbf{x}(t); \boldsymbol{\mu})$;
4: let $t = t + 1$;
5: **end while**

Implementation interpretations. Adopting the LB Algorithm, each D2D link first acquires CSI of the link from its transmitter to the BS. This can be either estimated based on the downlink signal (e.g., in a TDD uplink/downlink configuration), or provided by the BS, which measures the uplink channel and sends the information to the D2D user (e.g., in a frequency-division duplexing (FDD) uplink/downlink configuration). Apart from uplink

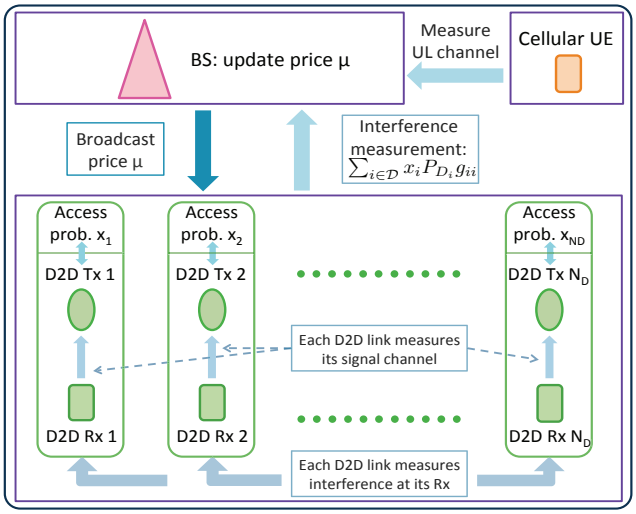


Figure 6.1: Illustration of the proposed algorithm. The arrows filled with dark color indicate the procedures requiring message exchange, while the arrows filled with light color indicate the procedures involving only local measurements. The lower part describes the LB Algorithm for the lower problem, while the upper part illustrates algorithms proposed for the upper problem.

CSI, each D2D link also measures the channel between the transmitter to its paired receiver. The frequency to update CSI depends on the channel variance. For example, in a slow mobility scenario, D2D links may just update the information once (at the beginning of each resource allocation period). At each iteration of the LB Algorithm, every D2D link measures the interference if accessing a RB. There is no additional message exchange in this step. Thus, the LB Algorithm only requires local information and reduces the overhead, as shown in Fig. 6.1.

Convergence analysis. To get the convergence criteria of the LB Algorithm, we first investigate some basic properties of function f_L. We as-

sume that there are finite number of D2D links. We call the set of D2D links with $x_i = 1$ saturated D2D links, denoted by $\mathcal{S} := \{i \in \mathcal{D} : x_i = 1\}$, and the set of D2D links with $x_i \in (0, 1)$ active D2D links, denoted by $\mathcal{A} := \{i \in \mathcal{D} : x_i \in (0, 1)\}$. Denoting $\mathbf{s} = [s_1, \cdots, s_{N_D}]^T$, we have $\mathbf{s} = \mathbf{G}\mathbf{x}$, where $\mathbf{G}$ is an $N_D \times N_D$ matrix with zero diagonal elements and (i, j)th element (with $i \neq j$) being $P_{D_j} h_{ji}$.

Proposition 6.3. *The best-response function f_L has the following properties:*

1. *f_L is a continuous mapping from $\mathcal{X}$ to $\mathcal{X}$.*

2. *f_L is piecewise affine, which means that f_L has the following two properties:*

   (a) *The domain of f_L can be partitioned into finitely many polyhedral regions, denoted by $\mathcal{P}_1, \cdots, \mathcal{P}_d$, which are determined by $\mathcal{A}$ and $\mathcal{S}$;*

   (b) *On the polyhedron $\mathcal{P}_n$ defined by $\mathcal{A}^{(n)}$ and $\mathcal{S}^{(n)}$, we have*

   $$f_L(\mathbf{x}) = \mathbf{M}^{(n)}\mathbf{x} + \mathbf{b}^{(n)},$$

   where $\mathbf{b}^{(n)}$ is a constant vector, and $\mathbf{M}^{(n)} = \mathbf{B}^{(n)}\mathbf{G}$ with $\mathbf{B}^{(n)}$ being a diagonal matrix, which has $\left[\mathbf{B}_i^{(n)}\right]_{kl} = -\frac{1}{P_{D_i} h_{ii}}$ if $k = l, i \in \mathcal{A}^{(n)}$, and $\left[\mathbf{B}_i^{(n)}\right]_{kl} = 0$ otherwise.

Proof. See Appendix 6.9.1. □

We assume that the resource allocation is carried out well during the channel coherence time, and thus channel can be regarded as static during

resource allocation updates. We leave the stochastic channel analysis as future work. Defining the matrix norm of a matrix $\mathbf{M}$ induced by the vector norm $\|\cdot\|$ as $\|\mathbf{M}\| := \max\{\|\mathbf{Mx}\| : \|\mathbf{x}\| = 1\}$ [174], a sufficient condition for the convergence of the proposed algorithm with general matrix norms can be found as follows, leveraging the techniques used in Theorem 7 in [174].

Theorem 6.1. *If* $\|\mathbf{M}_n\| < 1$*, we have*

1. *the synchronous iterative algorithm converges for any initial resource allocation;*

2. *there is a unique fixed point* $\mathbf{x}^*$*;*

3. $\|\mathbf{x}(t) - \mathbf{x}^*\| \leq \eta^t \|\mathbf{x}(0) - \mathbf{x}^*\|$*, where* $\eta = \max_n \|\mathbf{M}_n\|$*. The upper bound of the convergence rate is* η*.*

Proof. If f_L is a contraction mapping, then global stability follows from the Banach Fixed Point Theorem (see, e.g.[178]). The proof of contraction mapping is similar to the proof of Theorem 7 in [174], and thus we ignore the details. Given that f_L is a contraction mapping, we have $|f_L(\mathbf{x}') - f_L(\mathbf{x})\| \leq \eta\|\mathbf{x} - \mathbf{x}'\|$. The rate of convergence for a sequence $\{x_n\}$ converging to L is defined as the $\lim_{n\to\infty} \frac{|x_{n+1}-L|}{|x_n-L|}$. Observing the above inequality, we conclude that the convergence rate of the BR Algorithm is upper bounded by η. □

The number of polyhedral regions that partition the domain of f_L is $O(3^{N_D})$, which is very large, and thus it is impractical to check the conditions

in Theorem 6.1 directly for all regions. We further provide sufficient conditions in Proposition 6.4 that are easy to apply.

Proposition 6.4. *If the matrix* $\mathbf{G}$ *satisfies* $\|G\| \leq \min_{i,k} P_{D_i} h_{ii}^{(k)}$, *then the algorithm converges to the unique fixed point regardless of the initial point.*

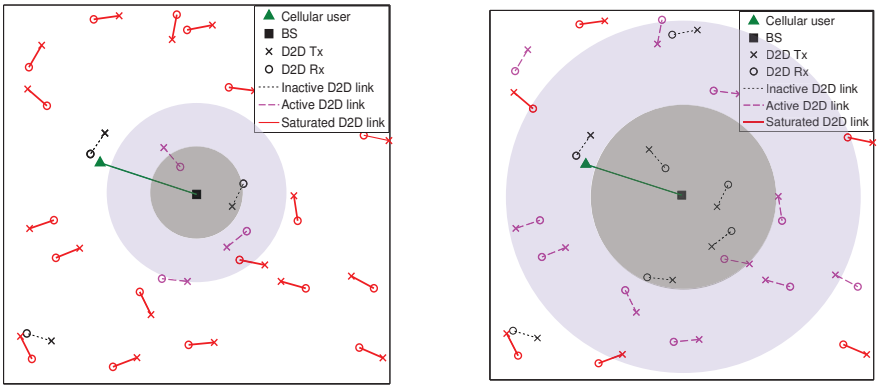
Proof. According to Prop. 6.3, we have $\mathbf{M}_n = \mathbf{B}^{(n)}\mathbf{G}$. To make f_L a contraction mapping, we have to satisfy $\|\mathbf{B}^{(n)}\mathbf{G}\| \leq 1$. By the property of matrix norm that $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$, we obtain a sufficient condition that is $\|\mathbf{B}^{(n)}\| \cdot \|\mathbf{G}\| \leq 1$. Matrix $\mathbf{B}^{(n)}$ is a diagonal matrix, whose norm is $\|\mathbf{B}^{(n)}\| \leq \max_i \frac{1}{P_{D_i} h_{ii}}$, $\forall n$. Then we can get one sufficient condition as $\|G\| \leq \left(\max_{i,k} \frac{1}{P_{D_i} h_{ii}^{(k)}}\right)^{-1} = \min_{i,k} P_{D_i} h_{ii}^{(k)}$. □

Design interpretations. The above result is true for any general l_p norm with $p \geq 1$. As in [174], we apply it to some special matrix norms and give the corresponding interpretations as follows.

Example 1 (l_1 norm). We have $\|\mathbf{G}\|_1 = \max\{\sum_{i=1} |s_i|\}$. This implies that a sufficient condition for the convergence of the LB Algorithm is that no D2D transmitter causes very strong interference to other D2D links.

Example 2 (l_∞ norm). We have $\|\mathbf{G}\|_\infty = \max\{\max_i |s_i|\}$. This implies that a sufficient condition for the convergence is that no D2D receiver suffers excessive interference.

We show examples of D2D access probabilities obtained by the BR Algorithm versus different μ in Fig. 6.2. We can observe that the BR Algorithm



(a) $\mu = 80$dB (b) $\mu = 90$dB

Figure 6.2: The access probabilities of D2D links vs. μ. The areas in the dark shade show the locations of silent D2D links with $x_i = 0$. The light shaded areas show the locations of active D2D links (i.e., $x_i \in (0, 1)$). The remaining parts show the locations of saturated D2D links (i.e., $x_i = 1$).

discourages D2D links whose transmitters are near the BS from accessing the RB. In addition, the BR Algorithm takes into account the SINR of D2D links, and encourages a D2D link far from the BS to keep silent if there are many D2D links nearby, to decrease the interference in D2D networks. Comparing the two subfigures, we conclude that less D2D links would be active as μ increases. This suggests the potential effectiveness of μ, which is investigated in the following section.

6.6 Upper Problem: Network's Pricing Mechanism

It is difficult to analyze the upper problem by plugging (6.6) directly into (6.4), due to the complex term in the denominator. Simulations in Section 6.7 show that the performance of the LB Algorithm for (6.7) is very close to the

performance of the BR Algorithm for the original problem (6.5). This suggests to approximate the solution of (6.5) to the solution of the lower bound problem (6.8). By this approximation, we propose an algorithm leveraging techniques from LCP [179]. We further propose a bisection algorithm, which has low overhead and can be applied to the original two-stage problem (with the lower problem (6.5)). In addition, motivated by the results of the lower problem (see e.g., Fig. 6.2), we propose a simple greedy heuristic algorithm, which performs well for networks with high interference tolerance level.

6.6.1 An Equivalent Upper Problem

According to Lemma 1 in [167], the upper problem is equivalent to the following problem:

$$\begin{aligned} &\max_{\mu \geq 0} \ \min \left\{ \mu \sum_{i \in \mathcal{D}} x_i P_{D_i} g_{ii}, \ \mu Q \right\} \\ &\text{s.t. } x_i = x_i^*, \end{aligned} \tag{6.9}$$

where x_i^* is given by (6.8). For simplicity, we use the notation $0 \leq a \perp b \geq 0$ to represent the complementarity condition of a and b, i.e., $ab = 0$ and $a, b \geq 0$ [179]. Letting $I_{Ci} = \sum_{j \in \mathcal{C}} P_{C_j} h_{ji} + W_i$, and λ_i be the Lagrange multiplier to relax the constraint $x_i \leq 1$, we have the following lemma.

Lemma 6.2. *Denoting* $t_i = \frac{w_i P_{D_i} h_{ii}}{u P_{D_i} g_{ii} \left(u P_{D_i} g_{ii} + \lambda_i \right)}$, *(6.8) is equivalent to the following parametric LCP with variables* (x_i, t_i) *and parameter* μ *[167, 179]:*

$$\begin{aligned} &0 \leq x_i \perp \left(t_i - \frac{w_i h_{ii}}{\mu g_{ii}} + \sum_{j \in \mathcal{D}} P_{D_j} h_{ji} x_j + I_{Ci} \right) \geq 0, \\ &0 \leq t_i \perp (1 - x_i) \geq 0. \end{aligned} \tag{6.10}$$

Proof. Details can be found in [167]. Key steps are to multiply (6.8) by $\frac{\sum_{j\in\mathcal{D}} P_{D_j} h_{ji} x_j + I_{Ci}}{\mu P_{D_i} g_{ii} + \lambda_i}$ and change variables t_i. □

In the following, we explore properties of the objective function in (6.9), which provides clues to design efficient algorithms for the upper problem.

6.6.2 Algorithm Design for the Upper Problem

In this chapter, we use the symmetric parametric principle pivoting algorithm (SPPP) – a classical algorithm for parametric LCP [179] – to find the optimal μ and its corresponding feasible solutions x_i in (6.10). We write (6.10) in matrix form as $\mathbf{0} \leq \mathbf{y} \perp \mathbf{A}\mathbf{y} + \mathbf{q} + \nu\mathbf{d} \geq 0$, where

$$\mathbf{y} = [x_1, \ldots, x_{N_D}, t_1, \ldots, t_{N_D}]^T,$$

$\mathbf{A} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix}$, with $\mathbf{A}_0$ is a matrix with (i,j)th element $P_{D_j} h_{ji}$, $\nu = \frac{1}{\mu}$, $\mathbf{q} = [I_{C1} \ldots, I_{CN_D}, 1, \ldots, 1]^T$, and $\mathbf{d} = [-\frac{w_1 h_{11}}{g_{11}}, \ldots, -\frac{w_{N_D} h_{N_D N_D}}{g_{N_D N_D}}, 0, \ldots, 0]^T$. Note that $\mu \geq 0$ implies that $\nu \geq 0$. We set a upper limit for ν, denoted by $\bar{\nu} < \infty$, which is a sufficient large real number.

The SPPP is given by Algorithm 2. Since the SPPP Algorithm requires CSI between each transmitter and receiver (i.e., matrix $\mathbf{A}$), which may cause high overhead, the result of SPPP can be used as a performance benchmark, and another low-overhead algorithm is desirable. To propose such algorithms, we first explore the properties of the objective function in (6.9), which is denoted by $U_c := \min\{U_{c1}, U_{c2}\}$, where $U_{c1} = \mu \sum_{i\in\mathcal{D}} x_i P_{D_i} g_{ii}$ and $U_{c2} = \mu Q$. Function U_{c2} is a linear increasing function of μ, while U_{c1} is more complicated

Algorithm 2 SPPP-based algortihm [179]

1: Initialization: $\tau = 0$, $\mu^* = 0$, $U^* = 0$, $\mathbf{q}(\tau) = \mathbf{q}$, $\mathbf{d}(\tau) = \mathbf{d}$, $\mathbf{A}(\tau) = \mathbf{A}$, $\nu(\tau) = 0$, $\mathbf{y}(\tau) = \mathbf{1}$, and
$$\mathbf{z}(\tau) = \mathbf{q} + \nu(\tau)\mathbf{d} + \mathbf{A}\mathbf{y}(\tau);$$
▷ **comment**: critical value
2: Determine the next critical value of μ:
$$\nu(\tau+1) = \min\left\{\min_i\left\{-\frac{q_i(\tau)}{d_i(\tau)} : d_i(\tau) < 0\right\}, \bar{\nu}\right\};$$
3: Set $(\mathbf{z}(\tau+1), \mathbf{y}(\tau+1)) = (\mathbf{q}(\tau) + \nu\mathbf{d}(\tau), 0)$ for all $\nu \in [\nu(\tau), \nu(\tau+1)]$;
4: **if** $\nu(\tau+1) = \bar{\nu}$ **then**
5: stop;
6: **else**
7: let $r = \arg\min_i \left\{-\frac{q_i(\tau)}{d_i(\tau)} : d_i(\tau) < 0\right\}$;
8: **end if**
9: The new critical value of ν is $\nu(\tau+1) = -q_r(\tau)/d_r(\tau)$, and thus $\mu(\tau+1) = (\nu(\tau+1))^{-1}$;
▷ **comment**: pivoting
10: **if** $A_{rr}(\tau) > 0$ **then**
11: pivot $< z_r(\tau), y_r(\tau) >$;
12: let $z_r(\tau+1) = y_r(\tau)$, $y_r(\tau+1) = z_r(\tau)$;
13: let $z_i(\tau+1) = z_i(\tau)$, $y_i(\tau+1) = y_i(\tau)$, for $i \neq r$;
14: let $\tau = \tau + 1$, and return to Step 2;
15: **else if** $A_r r(\tau) = 0$ **then**
16: use $y_r(\tau)$ as a driving variable and determin the basic blocking variable $z_s(\tau)$;
17: pivot $< z_s(\tau), y_r(\tau) >$, $< z_r(\tau), y_s(\tau) >$;
18: let $z_s(\tau+1) = y_r(\tau)$, $y_s(\tau+1) = z_r(\tau)$, $z_r(\tau+1) = y_s(\tau)$, $y_r(\tau+1) = z_s(\tau)$;
19: let $z_i(\tau+1) = z_i(\tau)$, $y_i(\tau+1) = y_i(\tau)$, for $i \neq r, s$;
20: let $\tau = \tau + 1$, and return to Step 2;
21: **end if**
22: get $x_i(\tau+1)$ from $y_i(\tau+1)$;
23: let U be (6.9) at $\mu(\tau+1)$ and $x_i(\tau+1)$;
24: **if** $U > U^*$ **then**
25: let $U^* = U$, $\mu^* = \mu(\tau+1)$;
26: **end if**

since it involves (6.8). The properties of function U_{c1} are given by Proposition 6.5, leveraging the techniques in [167].

Proposition 6.5. *The function $U_{c1}(\mu)$ has the following properties:*

1. *U_{c1} is a continuous function of μ;*

2. *U_{c1} is piecewise affine;*

3. *If $\sum_{j\in\mathcal{D},j\neq i}\frac{h_{ij}}{h_{jj}}g_{jj} < g_{ii},\ \forall i\in\mathcal{D}$, and $\sum_{j\in\mathcal{S}}P_{D_j}\left(h_{ji}-\frac{h_{1i}}{g_{11}}g_{jj}\right)\geq 0,\ \forall i\in\mathcal{A}$, then U_{c1} is a non-increasing function.*

Proof. See Appendix 6.9.2. □

The sufficient conditions given in Proposition 6.5 to make U_{c1} non-decreasing essentially say that the interference among D2D links and the interference from saturated D2D links (i.e., $x_i = 1$) to the BS should be weak. Given by Proposition 6.5 that U_{c1} is piecewise affine, and U_{c2} is linear, the optimal μ^* must either be at a *break point* of U_{c1} – the discontinuous points in the derivative of U_{c1} – or at the intersection of U_{c1} and U_{c2} [167]. When U_{c1} is non-decreasing, the optimal μ^* must be at the intersection of U_{c1} and U_{c2}. This motivates the heuristic algorithm, given by Algorithm 3, which converges to an intersection point of U_{c1} and U_{c2} [167].

In the bisection algorithm, we let $\mu_{\max} < \infty$ be a sufficient large real number as an upper limit for μ. We consider non-trivial cases, where the interference from D2D to BSs, when all D2D links are active, is greater than

Algorithm 3 A bisection algorithm for finding optimal price μ^* at monotonic case

Initialization: given accuracy $\epsilon \geq 0$, let $\mu_u = \mu_{\max}$, and $\mu_l \geq 0$;
while $|\mu_u - \mu_l| \geq \epsilon$ **do**
    let $\mu_m = \frac{\mu_u + \mu_l}{2}$;
    get $x_i(\mu_m)$ by running the LB Algorithm;
    if $U_1(\mu_m, x_i(\mu_m)) \leq U_2(\mu_m, x_i(\mu_m))$ **then**
        $\mu_u = \mu_m$;
    else
        $\mu_l = \mu_m$;
    end if
end while
let $\mu^* = \mu_m$.

the interference tolerance level; otherwise, we just let all D2D links access the RB with probability one. Under this assumption, we have the following result.

Proposition 6.6. *The bisection algorithm always converges. In particular, the algorithm requires at most* $\log_2(\mu_{\max}/\epsilon)$ *iterations to converge.*

Proof. See Appendix 6.9.3. □

Note that the bisection algorithm also converges when we use the BR Algorithm instead of the LB Algorithm to solve the lower problem, due to that function f is continuous. Under the conditions given by Proposition 6.5, i.e., the interference among D2D links and the interference from saturated D2D links to BSs are weak, the bisection algorithm achieves the optimal μ^*. In other words, the optimal strategy in this case is to let the number of active D2D links as large as possible, until the total interference from D2D links reaches the tolerance level.

Implementation interpretations. Adopting the bisection algorithm, the BS first broadcasts a price, and then measures the aggregate interference at this price. If the interference is greater than the tolerance level, the BS increases the price; otherwise, the BS decreases the price. In fact, the behavior is consistent with the *law of supply and demand*: if the demand (the interference) exceeds the supply (the interference tolerance level), the price increases to make the RB less attractive. The algorithm can also be implemented adaptively. The network locally measures the total D2D interference, and increases (decreases) the price if the interference level is above (below) the predefined tolerance level Q, until the interference level reaches Q. Fig. 6.1 illustrates the structure of Algorithm 3, which shows that the signalling overhead is caused by the price broadcast and the channel measurements. The overhead due to price broadcast is proportional to the number of RBs, which is quite small. As for the channel measurements, the BS requires the channel information of cellular links, and each D2D link needs the CSI of the link between its transmitter to the paired receiver and of the link between the transmitter and the BS. Thus, the algorithm only requires local information and the overhead due to the channel measurements is proportional to the total number of cellular and D2D links, which is much lower than the overhead of centralized algorithms (e.g., the brute force approach or the SPPP Algorithm) that require global CSI. Note that the channel information updating frequency depends on the channel variance over time, which is quite low in a slow mobility environment (e.g., we may only measure channels once for each or several resource alloca-

tion time scales). Therefore, the required overhead is not significant compared to the potential advantages of our algorithm.

The computational complexity of Algorithm 3 is $O(N_D \log_2(\mu_{\max}/\epsilon) + N_D^2 T)$, where T is the number of required iterations for the LB Algorithm. The parameters T and $\log_2(\mu_{\max}/\epsilon)$ are generally much smaller than $N_x^{N_D}$ as illustrated in Section 6.7, where T and $\log_2(\mu_{\max}/\epsilon)$ are between 5 and 10, while $N_x^{N_D}$ is 10^{10}. Thus, the complexity of Algorithm 3 is much lower than the complexity of the centralized scheme, which is $O(N_x^{N_D} N_D^2)$.

Observing Fig. 6.2, D2D links mostly have larger access probabilities when they are far from the BS. This motivates another greedy heuristic algorithm – called the *IO Algorithm* (short for interference ordering), which needs no iteration. The D2D links are sorted by the interference caused to the BS in an ascending order, i.e., $P_{D1} g_{11} \leq P_{D2} g_{22} \leq \cdots \leq P_{DN_D} g_{N_D N_D}$. The BS lets $x_1 = 1, \ldots, x_n = 1$ and other D2D links be silent, where n satisfies $\sum_1^n P_{Di} g_{ii} \leq Q$ and $\sum_1^{n+1} P_{Di} g_{ii} > Q$. Adopting the IO Algorithm, the BS measures the uplink CSI from D2D transmitters, based on which the BS determines the access probabilities. Therefore, this algorithm has lower overhead than the bisection algorithm, and gets the solution more quickly, at the cost of overall performance, which is shown in the following section.

6.7 Performance Evaluation

We consider an uplink system with a hexagonal BS model. The main simulation parameters are listed as follows, unless otherwise specified. The

BS density is 1 per $\pi\,(500\text{m})^2$. The cellular UEs and D2D links are deployed according to two independent Poisson point processes with the same density 10 links per macrocell. We let the average length of D2D links be 80m. We assume the total bandwidth is 10MHz with 1MHz per subband. The transmitters adopt fractional power control, i.e., $P = \min\{P_{\max}, d^{\kappa\alpha}\}$, where $P_{\max}$ is the maximum transmit power, d is the distance of the link, κ is the compensation factor for path loss, and α is the path loss exponent. We let cellular UEs and D2D links have the same power control factor $\kappa = 0.75$. The maximum transmit powers of cellular UEs and D2D links over one subband are 200mW and 20mW, respectively, due to the fact that cellular UEs only access one subband while D2D links can access multiple subbands. Note that D2D links may not access all subbands, and thus we set a conservative maximum transmit power for D2D links. The noise power spectrum density is -174 dBm/Hz. Path loss exponents of UE-UE and UE-BS links are 4.37 and 3.76, respectively. We compare the performance of our proposed algorithms to the scenario where all D2D links are active, as well as the scheme where D2D links become silent when their transmitters are within a circle around their nearest BSs – called a *guard zone scheme*.

6.7.1 The Lower Problem: D2D Non-cooperative Game

In this section, we consider a single cell scenario, where the interference tolerance level is 5dB above the cellular signal. We investigate the performance of the BR Algorithm and the LB Algorithm. Note that the BR Algorithm pro-

vides the NE result, which may not be optimal. Due to the complexity to solve (6.1) via brute force search ($O(N_x^{N_D} N_D^2)$), we compare the gap between the NE and optimal results in a small network with three D2D links. The average total D2D rates obtained by brute force search and by the BR Algorithm are 3.362 bps/Hz and 3.355 bpz/Hz, respectively. Thus, we observe that the NE solution is near-optimal in small networks, which is mainly due to that the D2D links are active with probability close to one in most cases in the small network. On the other hand, the D2D links have fractional active probabilities in most cases in the large network, and thus the observation may be quite different in large networks. We leave the analysis of the gap between the NE and optimal solution of (6.5) in more general networks to future work. To compare the BR Algorithm and LB Algorithm, we consider a case with ten D2D links. Fig. 6.3 shows that the rate distributions using the BR Algorithm and the LB Algorithm are almost the same. This implies that we can use the solution of the LB Algorithm to approximate the solution of the BR Algorithm. Comparing to Figs. 6.6 and 6.7, we observe that the performance of different algorithms in single-cell networks is similar to the multi-cell networks. Therefore, more discussion is left to the following subsection.

6.7.2 The Upper Problem: Network Pricing Mechanism

In this section, we consider an asynchronous multi-cell network, where each cell allocates resources independently. We let the interference tolerance level be the same as the received signal of cellular link (i.e., the interference

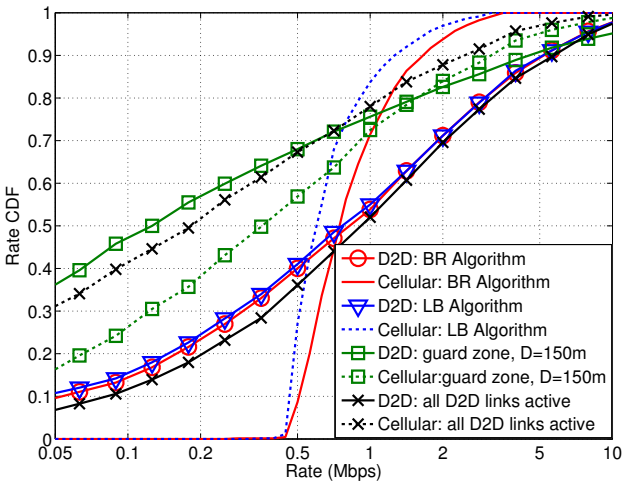


Figure 6.3: The rate distributions of D2D and cellular links using different algorithms in a single-cell network.

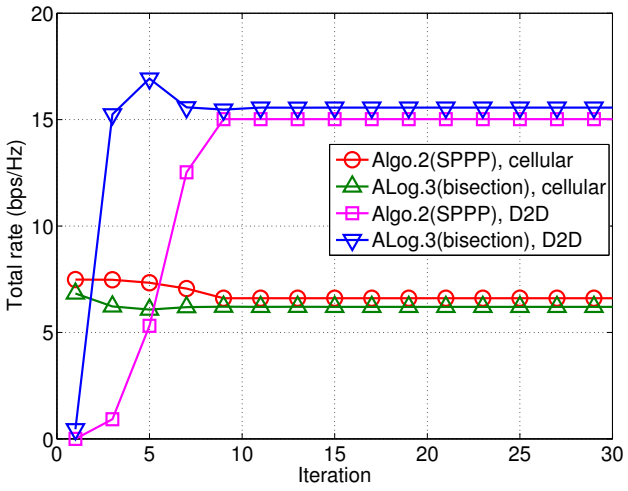


Figure 6.4: The convergence of different algorithms.

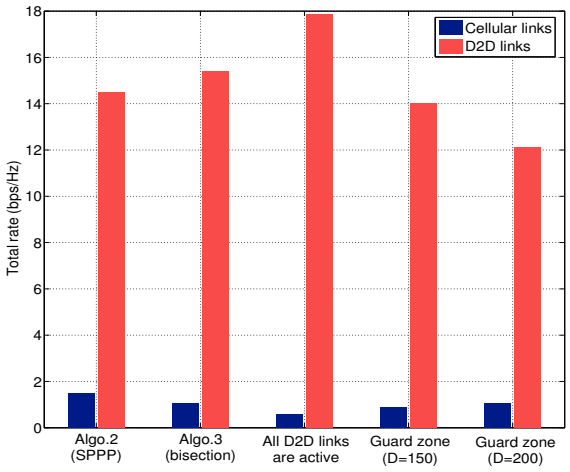


Figure 6.5: The total rates of cellular and D2D links using different approaches.

level normalized by the cellular signal is 0dB). From simulation results, the number of iterations required for the convergence of the LB Algorithm is about 4-8, which is quite small. Fig. 6.4 shows the convergence of the SPPP and bisection algorithms. Both algorithms converge quickly. While SPPP provides a larger cellular rate, it converges more slowly than the bisection algorithm. The quick convergence of LB Algorithm and bisection algorithm implies that the complexity of the proposed scheme $O(N_D \log_2(\mu_{\max}/\epsilon) + N_D^2 K)$ is much lower than the complexity of the centralized scheme $O(N_x^{N_D} N_D^2)$, where $K \in [4, 8]$ and $\log_2(\mu_{\max}/\epsilon) \in [5, 10]$, while $N_x \geq 2$ generally and thus $N_x^{N_D} \geq 2^{10}$ in our setup.

In Fig. 6.5, we compare the rates of D2D and cellular links using different algorithms. The SPPP and bisection algorithms provide larger D2D and/or cellular rates than the guard zone schemes. If there is no interfer-

ence management, the rate of cellular UE is very small (see "All D2D active" in the figure). Adopting the proposed algorithms, the cellular links can get much better performance (total cellular rate increasing from 0.61 to about 1.07 bps/Hz), at the cost of less total throughput (about 12% loss in our setup). We observe that SPPP provides a slightly larger rate for cellular links than the bisection algorithm. This implies that in some cases, the function U_{c1} is non-monotonic and the optimal μ is not at the intersection of U_{c1} and U_{c2}. However, in general, the gap between the bisection algorithm and the SPPP algorithm is small regardless of the monotonicity of U_{c1}. The average total rate in conventional networks, where potential D2D links operate only in cellular mode, is 2.4 bps/Hz in our setup. Defining the rate gain by the increased total rate divided by the average rate in conventional networks, we conclude that allowing D2D links and using proposed algorithms achieves a very large rate gain compared to conventional networks (about 5x in our simulation setup), and meanwhile keeps the performance of cellular UEs at an acceptable level (with average rate per cellular link being 1.07 bps/Hz). Note that the rate gain depends on various system parameters, such as average D2D link length and the amount of D2D traffic.

The rate distributions of cellular and D2D links are shown in Figs. 6.6 and 6.7, respectively. Fig. 6.6 shows that proposed algorithms can effectively protect the cellular performance. Comparing to Fig. 6.3, where the average cellular rate is about 1.5 bps/Hz with the normalized interference tolerance level being 5dB, we can conclude that a lower normalized interference tolerance

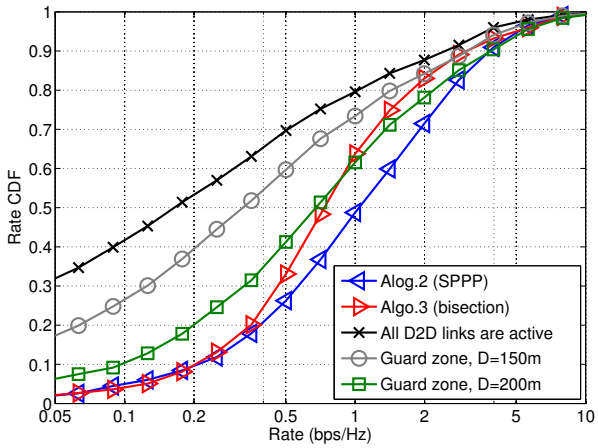


Figure 6.6: The rate distribution of cellular links using different approaches.

level (0dB) is needed in the multi-cell scenario. Also, the rate of cellular links has a larger range than the single cell scenario (i.e., the variance is larger). One possible reason is that there may be some nearby interfering D2D links and cellular UEs in neighboring cells. Though the D2D links have large rates without interference management, they hurt cellular links a lot. Adopting the guard zone scheme, cellular links can be protected, at the cost of the degradation of D2D throughput. Moreover, it is difficult to develop a tractable framework to study the guard zone scheme, and thus difficult to find the optimal distance threshold analytically. Therefore, the SPPP and bisection schemes are more preferable.

Note that the interference tolerance level can be tunable to maximize utility functions in terms of both the cellular and D2D links (e.g., the total rate in the hybrid network). We show the rates of cellular and D2D links versus the interference tolerance level numerically in Figs. 6.8 and 6.9, respectively.

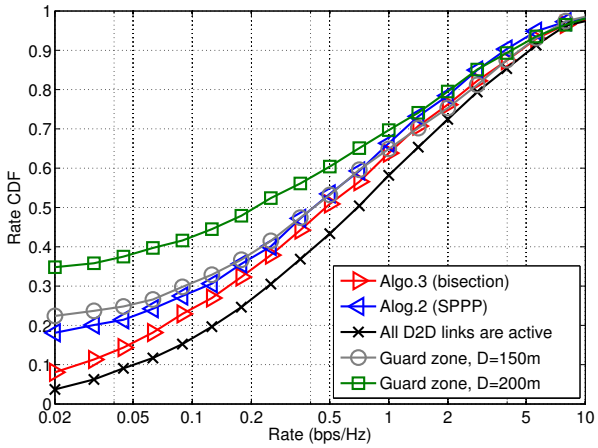


Figure 6.7: The rate distribution of D2D links using different approaches.

The analysis of optimal Q with respect to different utility functions is left to future work. Fig. 6.8 shows that as the interference tolerance level increases, the rate of cellular users decreases, because more D2D links are allowed to transmit. On the other hand, as Q increases, D2D links can access the RBs more aggressively and the total rate of D2D links increases. The IO Algorithm protects the performance of cellular links well. However, in a network with strict interference constraints, the total rate of D2D links using the IO Algorithm is less than the SPPP and bisection algorithms, which implies the importance to consider power control for D2D resource allocation, as well as the interference experienced at D2D receivers. From Fig. 6.9, we can conclude that although the IO Algorithm is very simple, it can only be applied to the cases with high interference tolerance (e.g., cases with normalized interference tolerance level larger than 0dB).

Figs. 6.10 and 6.11 show the rates of cellular links and D2D links versus

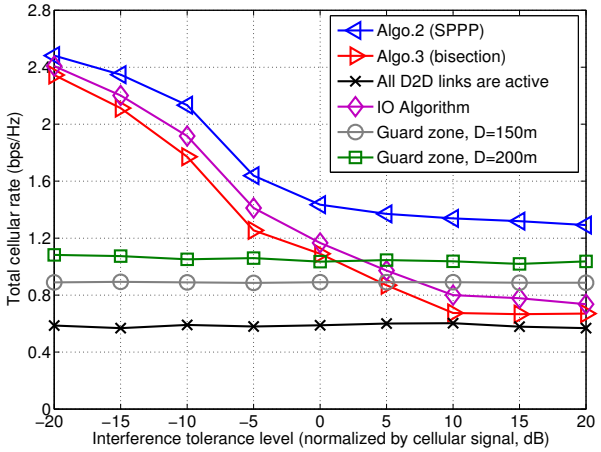


Figure 6.8: The rate of cellular links vs. the interference tolerance level. The normalized interference tolerance level means that the interference tolerance level Q is divided by the signal of the cellular link accessing the considered RB.

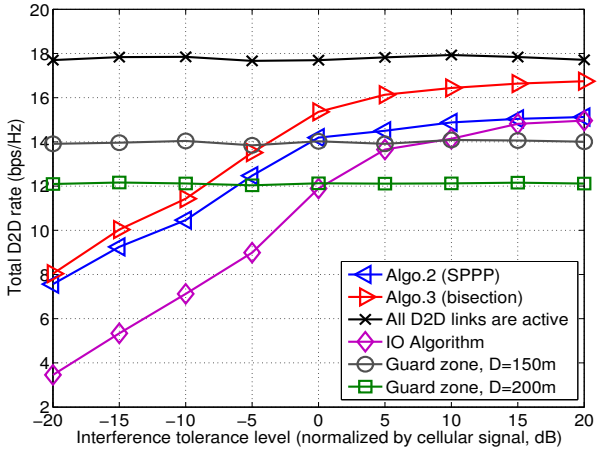


Figure 6.9: The total rate of D2D links vs. the interference tolerance level. The normalized interference tolerance level means that the interference tolerance level Q is divided by the signal of the cellular link accessing the considered RB.

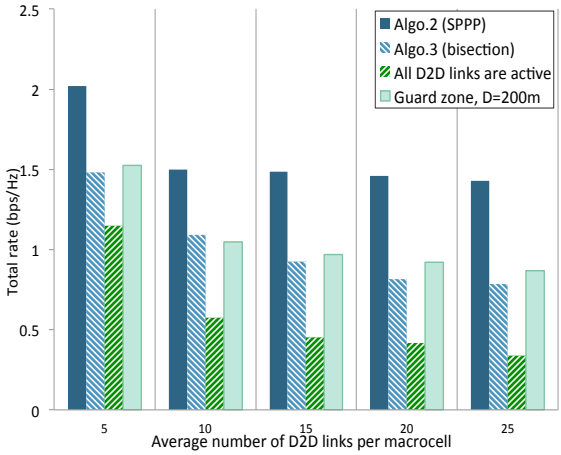


Figure 6.10: The total rate of cellular links vs. different D2D densities.

the D2D density, respectively. We ignore the guard zone scheme with radius 150m in these figures, since its performance is similar to the guard zone scheme with radius 200m. As shown in Fig. 6.11, the total rate of D2D links increases as D2D density increases, while the rate of cellular link decreases in Fig. 6.10. The decrease of cellular rate using the SPPP and bisection algorithms vanishes much more quickly than the scenario with all D2D links being active, which suggests the efficiency of the SPPP and bisection algorithms for protecting cellular transmissions. The figures also show that besides the interference tolerance level, the throughput gain of the proposed algorithms also highly depends on the density of D2D links.

6.8 Summary

This chapter presents a decentralized spectrum management for a shared network consisting of D2D and cellular links, aiming to maximize the total

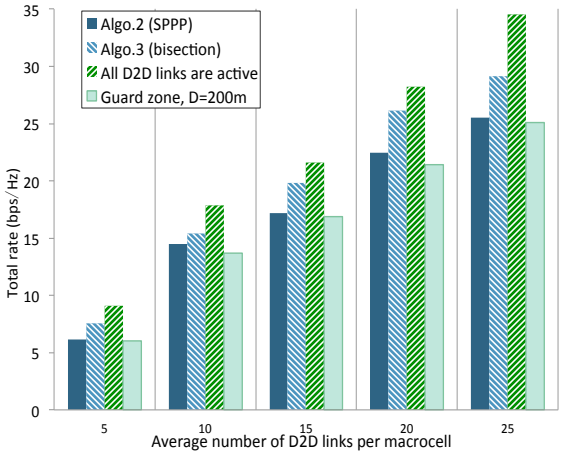


Figure 6.11: The total rate of D2D links vs. different D2D densities.

throughput of D2D links with an interference constraint for protecting cellular transmissions. We propose a low-complexity low-overhead distributed algorithm to update D2D access probabilities, and use the SPPP algorithm to get the optimal price for controlling the interference from cochannel D2D links. Though the SPPP Algorithm requires global CSI, it provides a benchmark for other algorithms. We further propose a low-overhead efficient heuristic algorithm based on the bisection method, which is shown to be convergent. Numerical results show that the heuristic algorithm has about the same performance as the SPPP algorithm, especially in the cases with low interference tolerance level. Another simple greedy algorithm is proposed and shown to perform well in scenario with high interference tolerance level. The proposed algorithms provide a large throughput gain with a performance guarantee of cellular links, compared to a conventional network with links operating only in cellular mode. Comparing to the cases without interference management (i.e.,

all D2D links are active), the average rate of cellular links improves significantly (e.g., average rate per cellular link increases from 0.61 to 1.07 bps/Hz in our setup). This implies that the proposed algorithms can efficiently manage the interference from D2D links to the cellular network. Future work could include investigation of more general utility functions incorporating both throughput and fairness, joint optimization of D2D mode selection, and consideration of a more flexible multiple-cell system.

6.9 Appendix

6.9.1 Proof of Proposition 6.3

1. We can complete the proof by applying the *Maximum Theorem* with $\Phi = U_i(\mathbf{x}_i; \mathbf{x}_{-i}, \boldsymbol{\mu})$ [174].

2. Let $\mathcal{A}$ and $\mathcal{S}$ be any two distinct subsets of set $\mathcal{D} = \{1, 2, \cdots, N_D\}$. We have

$$\begin{cases} a_i - s_i \leq 0, & \text{for } i \in \mathcal{D} \setminus (\mathcal{A} \cup \mathcal{S}), \\ 0 < a_i - s_i < P_{D_i} h_{ii}, & \text{for } i \in \mathcal{A}, \\ a_i - s_i \geq P_{D_i} h_{ii}, & \text{for } i \in \mathcal{S}. \end{cases} \tag{6.11}$$

   Given that the number of D2D links is finite, we have that the number of choices of disjoint $\mathcal{A}$ and $\mathcal{S}$ is finite. Therefore, the domain of f_L can be partitioned into finitely many polyhedra. With $\mathbf{s} = \mathbf{G}\mathbf{x}$, the inequalities (6.11) can be changed to inequalities in terms of $\mathbf{x}$, which defines a (possibly empty) polyhedron in $\mathcal{X}$. On the polyhedron $\mathcal{P}_n$, according to (6.8), we have $x_i = \frac{a_i - s_i}{P_{D_i} h_{ii}}$ for $i \in \mathcal{A}^{(n)}$, $x_i = 1$ for $i \in \mathcal{S}^{(n)}$ and $x_i = 0$ otherwise, which can

be expressed in a matrix form as $x_i = \mathbf{B}_i^{(n)} s_i + \mathbf{b}_i^{(n)}$, where $\mathbf{B}_i^{(n)}$ is defined in the Proposition 6.3. Combining with $\mathbf{s} = \mathbf{G}\mathbf{x}$, we complete the proof.

6.9.2 Proof of Proposition 6.5

1. We define a function $g : \mathbb{R}^+ \to \mathcal{X}$ as $g(\mu) = (g_1(\mu), \ldots, g_{N_D}(\mu))$, where $g_i(\mu) = x_i(\mu, \mathbf{x}_{-i}(0))$, $\mathbf{x}(0)$ is a given initial vector, and $x_i(\mu, \mathbf{x}_{-i}(0))$ is calculated according to (6.8) by fixing $\mathbf{x}_{-i} = \mathbf{x}_{-i}(0)$. Observing (6.8), we can see that $g_i(\mu)$ is a continuous function for a given $\mathbf{x}_{-i}$. According to properties of continuous functions (see, e.g., Theorem 4.10 in [180]), $g(\mu)$ is continuous due to the fact that each of the function $g_1(\mu), \ldots, g_{N_D}(\mu)$ is continuous. Proposition 6.3 shows that the best-response function $f_L : \mathcal{X} \to \mathcal{X}$ is continuous. Invoking properties of continuous functions again (see, e.g., Theorem 4.7 in [180]), we can conclude that $f_L(g(\mu)$ is a continuous mapping, which implies that the NE $x_i^*(\mu)$ is a continuous function of μ. Therefore, U_{c1} is also a continuous function of μ.

2. Recall $\mathcal{A}$ and $\mathcal{S}$ denote the sets of active D2D links and of saturated D2D links, respectively. Without loss of generality, let $\mathcal{A} = \{1, \ldots, n\}$ and $\mathcal{S} = \{n+1, \ldots, n+m\}$. We denote $\mathbf{x}_a = [x_1, \cdots, x_n]^T$, $\mathbf{x}_s = [x_{n+1}, \cdots, x_{n+m}]^T$ and $\mathbf{x}_0 = [x_{n+m+1}, \cdots, x_{N_D}]^T$. Let $\mathbf{H}_{aa}$ be a matrix with (i,j)th element being $\frac{P_{D_j} h_{ji}}{P_{D_i} h_{ii}}$ for $i, j \in \mathcal{A}$, $\mathbf{W}_a = \left[\frac{w_1}{P_{D_1} g_{11}}, \cdots, \frac{w_n}{P_{D_n} g_{nn}}\right]^T$ and $\mathbf{C}_a = \left[\frac{I_{C1}+I_{D_1}}{P_{D_1} h_{11}}, \cdots, \frac{I_{Cn}+I_{D_n}}{P_{D_n} h_{nn}}\right]^T$, where $I_{D_i} = \sum_{j \in \mathcal{S}} P_{D_j} h_{ji}$. According to (6.8), we have $\mathbf{x}_a = (\mathbf{H}_{aa})^{-1} \left[\frac{\mathbf{W}_a}{\mu} - \mathbf{C}_a\right]$, $\mathbf{x}_s = 1$ and $\mathbf{x}_0 = 0$. The domain of function U_{c1} can be divided into finite polyhedra according to differ-

ent $\mathcal{A}$ and $\mathcal{S}$. Denoting $\boldsymbol{\beta}_a = [P_{D_1} g_{11}, P_{D_2} g_{22}, \cdots, P_{D_n} g_{nn}]^T$ and $\boldsymbol{\beta}_s = \left[P_{D_{n+1}} g_{(n+1)(n+1)}, \cdots, P_{D_{n+m}} g_{(n+m)(n+m)}\right]^T$, in each polyhedron, we have

$$\begin{aligned} U_{c1}(\mu) &= \mu \sum_{i \in \mathcal{A}} x_i P_{D_i} g_{ii} + \mu \sum_{i \in \mathcal{S}} P_{D_i} g_{ii} \\ &= \mu \boldsymbol{\beta}_a^T (\mathbf{H}_{aa})^{-1} \left[\frac{\mathbf{W}_a}{\mu} - \mathbf{C}_a \right] + \mu \boldsymbol{\beta}_s^T \mathbf{1}^T \\ &= \boldsymbol{\beta}_a^T (\mathbf{H}_{aa})^{-1} \left[\mathbf{W}_a + \left(\mathbf{H}_{aa} \tilde{\boldsymbol{\beta}}_1 \left(\boldsymbol{\beta}_s^T \mathbf{1}^T \right) - \mathbf{C}_a \right) \mu \right], \end{aligned} \tag{6.12}$$

where $\tilde{\boldsymbol{\beta}}_1 = [\frac{1}{P_{D_1} g_{11}}, 0, \cdots, 0]$, and $\mathbf{1} = [1, 1, \cdots, 1]^T$. Therefore, U_{c1} is a linear function in each given polyhedron, and thus it is piecewise affine.

3. We use the same argument as the proof of Theorem 1 in [167]. The first condition is equivalent to that matrix $\mathbf{H}_{aa}^{(\boldsymbol{\beta})} := \mathrm{diag}(\boldsymbol{\beta}_a) \mathbf{H}_{aa} \left(\mathrm{diag}(\boldsymbol{\beta}_a)\right)^{-1}$ is strictly (column-wise) diagonally dominant. We have

$$\boldsymbol{\beta}_a^T \mathbf{H}_{aa}^{-1} \left(\mathbf{C}_a - \mathbf{H}_{aa} \tilde{\boldsymbol{\beta}}_1 \left(\boldsymbol{\beta}_s^T \mathbf{1}^T \right) \right) = \mathbf{1}^T \left(\mathbf{H}_{aa}^{(\boldsymbol{\beta})} \right)^{-1} \mathbf{C}_a^{(\boldsymbol{\beta})}, \tag{6.13}$$

where $\mathbf{C}_a^{(\boldsymbol{\beta})} = \mathrm{diag}(\boldsymbol{\beta}_a) \left(\mathbf{C}_a - \mathbf{H}_{aa} \tilde{\boldsymbol{\beta}}_1 \left(\boldsymbol{\beta}_s^T \mathbf{1}^T \right) \right)$. To show that U_{c1} is non-increasing, we need to show that (6.13) is non-negative. The first condition is a sufficient condition for $\mathbf{1}^T \left(\mathbf{H}_{aa}^{(\boldsymbol{\beta})} \right)^{-1} \geq 0$. Similar proof can be found in [167], and we ignore the details. The remaining proof is to show $\mathbf{C}_a - \mathbf{H}_{aa} \tilde{\boldsymbol{\beta}}_1 \left(\boldsymbol{\beta}_s^T \mathbf{1}^T \right) \geq 0$. The ith element of the left term of the above inequality is $\frac{1}{P_{D_i} h_{ii}} \left(I_{C_i} + \sum_{j \in \mathcal{S}} P_{D_j} h_{ji} - \frac{h_{1i}}{g_{11}} \sum_{j \in \mathcal{S}} P_{D_j} g_{jj} \right)$, which implies that $\sum_{j \in \mathcal{S}} P_{D_j} \left(h_{ji} - \frac{h_{1i}}{g_{11}} g_{jj} \right) \geq 0, \ \forall i \in \mathcal{A}$ is a sufficient condition to make $\mathbf{C}_a - \mathbf{H}_{aa} \tilde{\boldsymbol{\beta}}_1 \left(\boldsymbol{\beta}_s^T \mathbf{1}^T \right) \geq 0$. Combining with $\mathbf{1}^T \left(\mathbf{H}_{aa}^{(\boldsymbol{\beta})} \right)^{-1} \geq 0$, we can conclude that (6.13) is non-negative, and thus U_{c1} is non-increasing.

6.9.3 Proof of Proposition 6.6

We first show that the intersection of U_{c1} and U_{c2} on $[0, \mu_{\max}]$ is non-empty, and then show that the bisection algorithm converges to one of the intersection points. Prop. 6.5 shows that U_{c1} is a continuous function of μ. It is easy to observe that U_{c2} is also a continuous function of μ. Therefore, $U_{c1} - U_{c2}$ is a continuous function of μ. Recalling the assumption that when all D2D links are active, the interference from D2D to BSs is greater than the interference tolerance level, we have $U_{c1} - U_{c2} > 0$ when $\mu = 0$. On the other hand, when $\mu = \mu_{\max}$, we have $U_{c1} - U_{c2} < 0$. According to the intermediate value theorem, we can conclude that there is some number $\mu \in [0, \mu_{\max}]$ such that $U_{c1}(\mu) - U_{c2}(\mu) = 0$. In other words, there is at least one intersection point between U_{c1} and U_{c2} on $[0, \mu_{\max}]$.

Adopting the bisection algorithm, the interval is divided into two halves at each iteration. The interval at iteration t is denoted by $L(t) = [a_t, b_t]$, where $a_0 = 0$ and $b_0 = \mu_{\max}$. According to procedures of the bisection algorithm, we have $U_{c1}(a_t) \geq U_{c2}(a_t)$ and $U_{c1}(b_t) \leq U_{c2}(b_t)$ at each iteration t. Similar to the proof of the existence of intersection points on $[0, \mu_{\max}]$, we can show that there is at least one intersection point between U_{c1} and U_{c2} on $L(t)$. Therefore, the bisection algorithm preserves the existence of intersection points in current interval. The length of interval $L(t)$ has $|L(t)| = |L(t-1)|/2 = \cdots = \mu_{\max}/2^t$. It must stop when $|L(t)| \leq \epsilon$, which implies that the algorithm converges, and the maximum number of iteration for convergence, denoted by T, satisfies $\mu_{\max}/2^T = \epsilon$, i.e., $T = \log_2(\mu_{\max}/\epsilon)$.

Chapter 7

Conclusions

7.1 Summary

Small BSs and D2D communication are emerging important technology components for cellular networks to meet the drastic rise in wireless traffic demand. An immediate effect of the increasing network heterogeneity is the obsolescence of conventional resource allocation schemes, particularly the user association (D2D mode selection) and interference management schemes. The key challenges in the design of user association and interference management in such heterogeneous and irregular networks include the high computational complexity to solve the massive combinatorial user association problem over SINRs of all users and loads of all BSs, the coupled relationship between user association, scheduling and interference management, and the low-complexity low-overhead requirement for algorithms in implementation. In this dissertation, we tackled these challenges with novel models and fundamental analysis leveraging techniques from optimization and stochastic geometry. Chapters 2, 3 and 4 focus on HetNets consisting of macro BSs and small cells, while Chapters 5 and 6 investigate the cellular networks integrating D2D communication. The main contributions are summarized as follows.

For the networks with small BSs, Chapter 2 proposes a novel user association scheme that achieves load balancing through a network-wide utility maximization problem. We propose a low-complexity distributed algorithm that approaches the optimal performance. The optimal performance is then adopted as a benchmark to find the biasing factors in CRE. We show that the simple CRE with identical biasing factor per tier provides near-optimal performance, if the biasing factors are carefully designed. The load-aware association significantly benefits the rate distribution.

Chapter 3 then extends the above framework to jointly optimize the user association and RB blanking, in the effort to further improve the cell-edge performance. We propose to relax the unique association, which converts the problem to a convex optimization. We show both theoretically and through simulation that the optimal association of the relaxed problem is still mostly unique. There is a significant difference in user association on normal and blank RBs in simulation, which implies the importance of jointly investigation of user association and interference management. Simulation shows that the RB blanking can further improve the rate of cell-edge users.

Then we extend the baseline single-antenna model to multiple-antenna transmission in Chapter 4. We focus on the case where BSs operate in the massive MIMO regime. We formulate a utility maximization problem to jointly optimize user association and interference management, where we focus on two interference management techniques – RB blanking and JT. We propose an efficient algorithm approaching optimal solutions, which can be implemented

in a partially distributed manner with low overhead. We further propose a simple scheduling scheme yielding near-optimal resource allocations. Both RB blanking and JT provide a great rate gain (1.6x for blanking and 2.2x for JT versus optimal user association without interference management) for the cell-edge users.

For the D2D enhanced cellular networks, Chapter 5 presents a tractable model to investigate the resource allocation between D2D and cellular networks, as well as the mode selection of potential D2D links based on an Aloha-type time hopping scheme. We provide analytical SINR distribution and average rate expressions, that are applied to efficiently optimize the design principles such as the resource partition ratio and time-frequency hopping probabilities. With an appropriate resource partition, we observe that the dedicated method has a larger overall rate than the shared method in the downlink. D2D links would access frequency bands as many as needed and all potential D2D links are in D2D mode to maximize the total rate in fully loaded networks. The result can serve as an optimized lower bound for networks with more sophisticated D2D resource allocations.

Chapter 6 proposes a more dynamic resource allocation algorithm for D2D links that share uplink cellular resources. We propose a low-complexity low-overhead distributed algorithm to maximize the network throughput with a performance guarantee for cellular links, whose key idea is a pricing mechanism. That is, the BSs adapt the price on each RB according to the aggregate interference from D2D links, and D2D links then determine whether to ac-

cess a RB or not based on the achievable rate and price on that RB. Results show that proposed algorithms provide a significant throughput gain, while maintaining the quality of cellular links at a predefined service level.

7.2 Future Directions

This dissertation has shown that load balancing with interference management is a key source of gain in HetNets. Despite the urgent need in the rethinking and re-investigation of metrics, intuitions and principles in designing HetNets, the load balancing problem is far from being fully understood. This dissertation concludes with some promising directions for future research.

Resource allocation for more dynamic cases. The considered models in this dissertation make several assumptions for tractability. For example, the system is fully loaded ("always on") and users are static over a sufficient large time window (e.g., the time scale of an association period). For dynamic networks (e.g., with dynamic traffic, time-variant channel and user mobility), resource allocation adapts to the dynamic variations may further improve the network performance, while directly applying the schemes proposed for static environment may limit or even kill the performance gain. For example, an optimal user association based on the static setting is to handover users to small cells as they enter the small cell area, and then back to macro BS as they leave the small cell. Such association scheme leads to very frequent handover for high-mobility users, which results in costly overhead and high power consumption. Thus, it may be preferable to associate the users with

high mobility to some suboptimal BSs, or use JT in MIMO systems to avoid handovers while providing good performance. Note that a related issue with user mobility is the difficulty in acquiring accurate CSI. That is, the algorithms need to adapt to the new network topology before the CSI is outdated. Thus, the algorithms designed for static case cannot be directly applied and it is of interest to explore how resource allocation such as user association and interference management can be tackled in dynamic systems.

Joint study of downlink and uplink association. The focus of load balancing study in networks with small cells in this dissertation is on the downlink. In conventional networks, the default association scheme in the uplink is typically the same as the association in downlink, since the coverage areas are almost the same among different macrocells. However, this is not the case in HetNets, since the BSs of different tiers have such widely divergent transmit powers. Rather, the downlink coverage area of macro BSs is much larger than that of smaller BSs. If we adopt the same association in uplink, the cell-edge macro-users will cause great interference to nearby users, especially for users which are associated to nearby small cells. Thus, it is necessary to jointly investigate downlink and uplink associations.

There are some recent interesting papers investigating the uplink association in HetNets. Papers [181, 182] study the uplink association using game theory techniques, while [183] proposes a heuristic algorithm providing the same association in both the downlink and uplink, aiming to maximize the downlink capacity and to minimize the uplink power. Different from the

aforementioned work, [184–187] validate our discussion that different associations in the uplink and downlink are desirable. Papers [184–186] use the stochastic geometry approach, where [184] shows that the minimum path loss association is optimal to maximize the uplink rate coverage, while [185, 186] focus on the total rate metric. Paper [187] presents system simulations that show a large uplink throughput gain with different uplink and downlink associations. In addition, [188] proposes a cooperative uplink reception scheme that allows the data to be decoded at users' best uplink reception nodes rather than the associated nodes. Overall, the optimal downlink and uplink associations are still far from being fully understood, and it is of interest to study the optimal downlink and uplink associations jointly with different interference management techniques in HetNets.

Integration of small BSs and D2D communication. This dissertation studies the load balancing in networks with small BSs and with D2D communication, respectively. It is interesting to extend the work to integrate both small BSs and D2D communication. In such HetNets, there are many resource allocation aspects that needed to be designed: user-BS association, D2D mode selection, resource allocation between D2D and cellular transmissions, and interference management for inter-cell interference as well as D2D-cellular interference. These coupled issues significantly complicates the load balancing problem. The interplay of small cell offloading and D2D offloading is an interesting open issue.

Bibliography

[1] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations research*, vol. 53, no. 1, pp. 12–25, Feb. 2005.

[2] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE Comm. Mag.*, vol. 50, no. 6, pp. 54–64, June 2012.

[3] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Comm. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[4] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: Old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, Apr. 2014.

[5] Ericsson, "Ericsson mobility report: On the pulse of the networked society," white paper, June 2014.

[6] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2014 - 2019," white paper, Feb. 2015.

[7] M. Dohler, R. W. Heath, A. Lozano, C. B. Papadias, and R. A. Valenzuela, "Is the PHY layer dead?" *IEEE Comm. Mag.*, vol. 49, no. 4, pp. 159–165, Apr. 2011.

[8] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Comm. Mag.*, vol. 51, no. 6, pp. 20–27, June 2013.

[9] Qualcomm, "LTE direct always-on device-to-device proximal discovery," *white paper*, Aug. 2014. Available: https://www.qualcomm.com/media/documents/files/lte-direct-always-on-device-to-device-proximal-discovery.pdf.

[10] Deutsche Telekom AG, Orange Silicon Valley, Qualcomm Technologies Incorporated, Tagged Incorporated, and Samsung Electronics, "LTE direct workshop white paper," May 2013. Available: https://www.qualcomm.com/media/documents/ files/lte-direct-whitepaper.pdf.

[11] Wi-Fi Alliance, "Wi-Fi peer-to-peer (P2P) technical specification," Feb. 2013.

[12] K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, and K. Hugl, "Device-to-Device communication as an underlay to LTE-Advanced networks," *IEEE Comm. Mag.*, vol. 47, no. 12, pp. 42–49, Dec. 2009.

[13] M. S. Corson, R. Laroia, J. Li, V. Park, T. Richardson, and G. Tsirtsis,

"Toward proximity-aware internetworking," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 26–33, Dec. 2010.

[14] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "FlashLinQ: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Trans. on Networking*, vol. 21, no. 4, pp. 1215–1228, Aug. 2013.

[15] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Turányi, "Design aspects of network assisted device-to-device communications," *IEEE Comm. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.

[16] M. Yavuz, F. Meshkati, S. Nanda, A. Pokhariyal, N. Johnson, B. Raghothaman, and A. Richardson, "Interference management and performance analysis of UMTS/HSPA+ femtocells," *IEEE Comm. Mag.*, vol. 47, no. 9, pp. 102–109, Sep. 2009.

[17] N. Saquib, E. Hossain, L. B. Le, and D. I. Kim, "Interference management in OFDMA femtocell networks: Issues and approaches," *IEEE Wireless Communications*, vol. 19, no. 3, pp. 86–95, June 2012.

[18] N. Himayat, S. Talwar, A. Rao, and R. Soni, "Interference management for 4G cellular standards [WIMAX/LTE UPDATE]," *IEEE Comm. Mag.*, vol. 48, no. 8, pp. 86–92, Aug. 2010.

[19] T. D. Novlan, R. K. Ganti, A. Ghosh, and J. G. Andrews, "Analytical evaluation of fractional frequency reuse for heterogeneous cellular networks," *IEEE Trans. on Communications*, vol. 60, no. 7, pp. 2029–2039, July 2012.

[20] M. Chiang, P. Hande, T. Lan, and C. W. Tan, "Power control in wireless cellular networks," *Foundations and Trends in Networking*, vol. 2, no. 4, pp. 381–533, Apr. 2008.

[21] V. Chandrasekhar, J. G. Andrews, T. Muharemovict, Z. Shen, and A. Gatherer, "Power control in two-tier femtocell networks," *IEEE Trans. on Wireless Communications*, vol. 8, no. 8, pp. 4316–4328, Aug. 2009.

[22] M. Rahman and H. Yanikomeroglu, "Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Trans. on Wireless Communications*, vol. 9, no. 4, pp. 1414–1425, Apr. 2010.

[23] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE Journal on Sel. Areas in Communications*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[24] 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 Communications Magazine*, vol. 18, no. 3, pp. 10–21, June 2011.

[25] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-Advanced: deployment scenarios and operational challenges," *IEEE Comm. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[26] D. Lopez-Perez, I. Guvenc, G. De La Roche, M. Kountouris, T. Q. Quek, and J. Zhang, *IEEE Wireless Communications*, vol. 18, no. 3, pp. 22–30, June 2011.

[27] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Comm. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[28] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. on Wireless Communications*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[29] P. Marques, J. Bastos, and A. Gameiro, "Opportunistic use of 3G uplink licensed bands," in *Proc., IEEE Intl. Conf. on Communications*, May 2008, pp. 3588–3592.

[30] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[31] T. Kahwa and N. Georganas, "A hybrid channel assignment scheme in large-scale, cellular-structured mobile communication systems," *IEEE Trans. on Communications*, vol. 26, no. 4, pp. 432–438, Apr. 1978.

[32] H. Jiang and S. Rappaport, "CBWL: A new channel assignment and sharing method for cellular communication systems," *IEEE Trans. on Veh. Technology*, vol. 43, no. 2, pp. 313–322, May 1994.

[33] S. K. Das, S. K. Sen, and R. Jayaram, "A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment," *Wireless Networks*, vol. 3, no. 5, pp. 333–347, Oct. 1997.

[34] ——, "A novel load balancing scheme for the tele-traffic hot spot problem in cellular networks," *Wireless Networks*, vol. 4, no. 4, pp. 325–340, July 1998.

[35] B. Eklundh, "Channel utilization and blocking probability in a cellular mobile telephone system with directed retry," *IEEE Trans. on Communications*, vol. 34, no. 4, pp. 329–337, Apr. 1986.

[36] X. Wu, B. Mukherjee, and S. H. G. Chan, "MACA-an efficient channel allocation scheme in cellular networks," in *Proc., IEEE Globecom*, vol. 3, Dec. 2000, pp. 1385–1389.

[37] D. Cavalcanti, D. Agrawal, C. Cordeiro, B. Xie, and A. Kumar, "Issues in integrating cellular networks WLANs, and MANETs: a futuristic

heterogeneous wireless network," *IEEE Wireless Communications Magazine*, vol. 12, no. 3, pp. 30–41, June 2005.

[38] E. Yanmaz and O. K. Tonguz, "Dynamic load balancing and sharing performance of integrated wireless networks," *IEEE Journal on Sel. Areas in Communications*, vol. 22, no. 5, pp. 862–872, June 2004.

[39] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc., IEEE INFOCOM*, vol. 1, Apr. 2003, pp. 786–796.

[40] Y. Bejerano and S. J. Han, "Cell breathing techniques for load balancing in wireless LANs," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 735–749, June 2009.

[41] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," *Wireless Networks*, vol. 14, pp. 103–120, Jan. 2008.

[42] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc., IEEE INFOCOM*, Apr. 2006, pp. 1–12.

[43] Y. Bejerano, S. J. Han, and L. Li, "Fairness and load balancing in wireless LANs using association control," *IEEE/ACM Trans. on Networking*, vol. 15, no. 3, pp. 560–573, June 2007.

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

[45] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. on Networking*, no. 99, pp. 1–14, June 2011.

[46] C. S. Chen, F. Baccelli, and L. Roullet, "Joint optimization of radio resources in small and macro cell networks," in *Proc., IEEE Veh. Technology Conf.*, May 2011, pp. 1–5.

[47] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *Proc., IEEE Intl. Conf. on Communications*, June 2012, pp. 2457–2461.

[48] 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. on Wireless Communications*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.

[49] S. Stanczak, M. Wiczanowski, and H. Boche, *Fundamentals of Resource Allocation in Wireless Networks: Theory and Algorithms.* Springer Verlag, 2009, vol. 3.

[50] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge university press, 2005.

[51] L. B. Jiang and S. C. Liew, "Proportional fairness in wireless LANs and ad hoc networks," in *Proc., IEEE Wireless Networking and Comm. Conf.*, vol. 3, Mar. 2005, pp. 1551–1556.

[52] S. Low and D. Lapsley, "Optimization flow control-I: basic algorithm and convergence," *IEEE/ACM Trans. on Networking*, vol. 7, no. 6, pp. 861–874, Dec. 1999.

[53] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ Pr, 2004.

[54] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

[55] D. P. Bertsekas, *Convex Optimization Theory*. Athena Scientific, 2009.

[56] B. Bjerke, "LTE-advanced and the evolution of LTE deployments," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 4–5, Oct. 2011.

[57] A. B. Saleh, O. Bulakci, S. Redana, B. Raaf, and J. Hamalainen, "Enhancing LTE-advanced relay deployments via biasing in cell selection and handover decision," in *Proc., IEEE PIMRC*, Sep. 2010, pp. 2277–2281.

[58] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "On/off macrocells and load balancing in heterogeneous cellular networks," in *Proc., IEEE Globecom*, Dec. 2013, pp. 3814–3819.

[59] S. Geirhofer and P. Gaal, "Coordinated multi point transmission in 3GPP LTE heterogeneous networks," in *IEEE Globecom Workshops*, Dec. 2012, pp. 608–612.

[60] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE Journal on Sel. Areas in Communications*, vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[61] I. Guvenc, M.-R. Jeong, I. Demirdogen, B. Kecicioglu, and F. Watanabe, "Range expansion and inter-cell interference coordination (icic) for picocell networks," in *Proc., IEEE Veh. Technology Conf.*, Sep. 2011, pp. 1–6.

[62] D. Lopez-Perez and H. Claussen, "Duty cycles and load balancing in hetnets with eICIC almost blank subframes," in *Proc., IEEE PIMRC*, Sep. 2013, pp. 173–178.

[63] M. Kamel and K. Elsayed, "ABSF offsetting and optimal resource partitioning for eICIC in LTE-Advanced: Proposal and analysis using a Nash bargaining approach," in *Proc., IEEE Intl. Conf. on Communications*, June 2013, pp. 6240–6244.

[64] L. Jiang and M. Lei, "Resource allocation for eICIC scheme in heterogeneous networks," in *Proc., IEEE PIMRC*, Sep. 2012, pp. 448–453.

[65] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[66] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. on Wireless Communications*, vol. 13, no. 2, pp. 888–901, Feb. 2014.

[67] M. Simsek, M. Bennis, and A. Czylwik, "Dynamic inter-cell interference coordination in HetNets: A reinforcement learning approach," in *Proc., IEEE Globecom*, Dec 2012, pp. 5446–5450.

[68] S. Lembo, P. Lunden, O. Tirkkonen, and K. Valkealahti, "Optimal muting ratio for enhanced inter-cell interference coordination (eICIC) in hetnets," in *Proc., IEEE Intl. Conf. on Communications*, June 2013, pp. 1145–1149.

[69] S. Vasudevan, R. Pupala, and K. Sivanesan, "Dynamic eICIC - a proactive strategy for improving spectral efficiencies of heterogeneous LTE cellular networks by leveraging user mobility and traffic dynamics," *IEEE Trans. on Wireless Communications*, vol. 12, no. 10, pp. 4956–4969, Oct. 2013.

[70] A. Bedekar and R. Agrawal, "Optimal muting and load balancing for eICIC," in *Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2013, pp. 280–287.

[71] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE hetnets," *IEEE/ACM Trans. on Networking*, vol. 22, no. 1, pp. 137–150, Feb. 2014.

[72] V. Gajic, J. Huang, and B. Rimoldi, "Competition of wireless providers for atomic users: Equilibrium and social optimality," in *Proc., Allerton Conf. on Comm., Control, and Computing*, Sep. 2009, pp. 1203–1210.

[73] Q. Ye, O. Y. Bursalioglu, and H. C. Papadopoulos., "Harmonized cellular and distributed massive MIMO: Load balancing and scheduling," in *submitted to Proc., IEEE Globecom*, Apr. 2015. Available at arXiv: http://arxiv.org/abs/1503.08227.

[74] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Comm. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[75] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[76] J. Hoydis, S. Ten Brink, M. Debbah *et al.*, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE Journal on Sel. Areas in Communications*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

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

[78] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Comm. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[79] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[80] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc., IEEE INFOCOM*, Apr. 2013, pp. 998–1006.

[81] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Trans. on Wireless Communications*, vol. 12, no. 3, pp. 1340–1351, Mar. 2013.

[82] N. Prasad, M. Arslan, and S. Rangarajan, "Enhanced interference management in heterogeneous cellular networks," in *Proc., IEEE Intl. Symposium on Information Theory*, June 2014, pp. 1603–1607.

[83] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks,"

submitted to IEEE Trans. Wireless Comm., Feb. 2015. Available at arXiv: http://arxiv.org/abs/1407.6731.

[84] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-Advanced [coordinated and distributed MIMO]," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 26–34, June 2010.

[85] J.-M. Moon and D.-H. Cho, "Inter-cluster interference management based on cell-clustering in network MIMO systems," in *Proc., IEEE Veh. Technology Conf.*, May 2011, pp. 1–6.

[86] P. Marsch and G. Fettweis, "Static clustering for cooperative multi-point (CoMP) in mobile communications," in *Proc., IEEE Intl. Conf. on Communications*, June 2011, pp. 1–6.

[87] J. Li, T. Svensson, C. Botella, T. Eriksson, X. Xu, and X. Chen, "Joint scheduling and power control in coordinated multi-point clusters," in *Proc., IEEE Veh. Technology Conf.*, Sep. 2011, pp. 1–5.

[88] Y. Cheng, S. Drewes, A. Philipp, and M. Pesavento, "Joint network optimization and beamforming for coordinated multi-point transmission using mixed integer programming," in *Proc., IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc. (ICASSP)*, Mar. 2012, pp. 3217–3220.

[89] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. on Wireless Communications*, vol. 12, no. 6, pp. 2762–2775, June 2013.

[90] Y. Du and G. de Veciana, ""wireless networks without edges": Dynamic radio resource clustering and user scheduling," in *Proc., IEEE INFOCOM*, Apr. 2014, pp. 1321–1329.

[91] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 118–127, June 2014.

[92] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Comm. Mag.*, vol. 41, no. 10, pp. 74–80, Oct. 2003.

[93] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: theoretical framework," *IEEE Trans. on Wireless Communications*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[94] ——, "Cross-layer optimization for OFDM wireless networks-part II: algorithm development," *IEEE Trans. on Wireless Communications*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[95] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE Journal on Sel. Areas in Communications*,

vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[96] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[97] L. Chen, S. H. Low, and J. C. Doyle, "Joint congestion control and media access control design for ad hoc wireless networks," in *Proc., IEEE INFOCOM*, vol. 3, Mar. 2005, pp. 2212–2222.

[98] ——, "Cross-layer design in multihop wireless networks," *Computer Networks*, vol. 55, no. 2, pp. 480–496, Feb. 2011.

[99] V. G. Subramanian, R. A. Berry, and R. Agrawal, "Joint scheduling and resource allocation in CDMA systems," *IEEE Trans. on Info. Theory*, vol. 56, no. 5, pp. 2416–2432, May 2010.

[100] J. Huang, V. G. Subramanian, R. Agrawal, and R. A. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Trans. on Wireless Communications*, vol. 8, no. 1, pp. 288–296, Jan. 2009.

[101] A. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination," in *Proc., IEEE INFOCOM*, Apr. 2009, pp. 1287–1295.

[102] I. Maric, B. Bostjancic, and A. Goldsmith, "Resource allocation for constrained backhaul in picocell networks," in *Information Theory and Applications Workshop (ITA)*, Feb. 2011, pp. 1–6.

[103] X. Wang and G. B. Giannakis, "Resource allocation for wireless multiuser OFDM networks," *IEEE Trans. on Info. Theory*, vol. 57, no. 7, pp. 4359–4372, July 2011.

[104] S. Rangan and R. Madan, "Belief propagation methods for intercell interference coordination in femtocell networks," *IEEE Journal on Sel. Areas in Communications*, vol. 30, no. 3, pp. 631–640, Apr. 2012.

[105] A. Bin Sediq, R. Schoenen, H. Yanikomeroglu, and G. Senarath, "Optimized distributed inter-cell interference coordination (ICIC) scheme using projected subgradient and network flow optimization," *IEEE Trans. on Communications*, vol. 63, no. 1, pp. 107–124, Jan. 2015.

[106] E. Leonardi, M. Mellia, F. Neri, and M. Ajmone Marsan, "On the stability of input-queued switches with speed-up," *IEEE/ACM Trans. on Networking*, vol. 9, no. 1, pp. 104–118, Feb. 2001.

[107] L. Chen, S. Low, M. Chiang, and J. Doyle, "Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks," in *Proc., IEEE INFOCOM*, Apr. 2006, pp. 1–13.

[108] I. C. Paschalidis, W. Lai, and D. Starobinski, "Asymptotically optimal transmission policies for large-scale low-power wireless sensor networks,"

IEEE/ACM Trans. on Networking, vol. 15, no. 1, pp. 105–118, Feb 2007.

[109] C. Joo, X. Lin, and N. Shroff, "Performance limits of greedy maximal matching in multi-hop wireless networks," in *IEEE Conference on Decision and Control*, Dec. 2007, pp. 1128–1133.

[110] G. R. Gupta, S. Sanghavi, and N. B. Shroff, "Node weighted scheduling," in *ACM SIGMETRICS*, vol. 37, no. 1, June 2009, pp. 97–108.

[111] A. Berger, J. Gross, and T. Harks, "The k-constrained bipartite matching problem: Approximation algorithms and applications to wireless networks," in *Proc., IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[112] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving large spectral efficiency with TDD and not-so-many base-station antennas," in *IEEE-APS Topical Conference on Antennas and Propagation in Wireless Communications (APWC)*, Sep. 2011, pp. 1346–1349.

[113] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. on Info. Theory*, vol. 56, no. 6, pp. 2845–2866, June 2010.

[114] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. on Wireless Communications*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[115] H. Huh, A. M. Tulino, and G. Caire, "Network MIMO with linear zero-forcing beamforming: Large system analysis, impact of channel estima-

tion, and reduced-complexity scheduling," *IEEE Trans. on Info. Theory*, vol. 58, no. 5, pp. 2911–2934, May 2012.

[116] Y.-G. Lim, C.-B. Chae, and G. Caire, "Performance analysis of massive MIMO for cell-boundary users," *arXiv preprint arXiv:1309.7817*, Sep. 2013.

[117] G.-C. Rota, "The number of partitions of a set," *The American Mathematical Monthly*, vol. 71, no. 5, pp. 498–504, May 1964.

[118] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2009. Available: http://cvxr.com/cvx/.

[119] H. Shirani-Mehr, G. Caire, and M. J. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Trans. on Communications*, vol. 58, no. 7, pp. 2055–2066, July 2010.

[120] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource allocation and cross layer control in wireless networks.* Now Publishers Inc, 2006.

[121] K.-K. Wong and Z. Pan, "Array gain and diversity order of multiuser MISO antenna systems," *International Journal of Wireless Information Networks*, vol. 15, no. 2, pp. 82–89, June 2008.

[122] A. Schrijver, *Theory of Linear and Integer Programming.* John Wiley & Sons, 1998.

[123] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Resource optimization in device-to-device cellular systems using time-frequency hopping," *IEEE Trans. on Wireless Communications*, vol. 13, no. 10, pp. 5467–5480, Oct. 2014.

[124] H. Nishiyama, M. Ito, and N. Kato, "Relay-by-smartphone: realizing multihop device-to-device communications," *IEEE Comm. Mag.*, vol. 52, no. 4, pp. 56–65, Apr. 2014.

[125] K. Huang, V. K. Lau, and Y. Chen, "Spectrum sharing between cellular and mobile ad hoc networks: transmission-capacity trade-off," *IEEE Journal on Sel. Areas in Communications*, vol. 27, no. 7, pp. 1256–1267, Sep. 2009.

[126] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Trans. on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[127] S. Xu, H. Wang, T. Chen, Q. Huang, and T. Peng, "Effective interference cancellation scheme for device-to-device communication underlaying cellular networks," in *Proc., IEEE Veh. Technology Conf.*, Sep. 2010, pp. 1–5.

[128] 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. on Wireless Communications*, vol. 10, no. 12, pp. 3995–4000, Dec. 2011.

[129] M. Zulhasnine, C. Huang, and A. Srinivasan, "Efficient resource allocation for device-to-device communication underlaying LTE network," in *IEEE Wireless Mobile Computing, Networking and Communications (WiMob)*, Oct. 2010, pp. 368–375.

[130] P. Phunchongharn, E. Hossain, and D. I. Kim, "Resource allocation for device-to-device communications underlaying LTE-Advanced networks," *IEEE Wireless Communications*, vol. 20, no. 4, pp. 91–100, Aug. 2013.

[131] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, and B. Jiao, "Interference-aware resource allocation for device-to-device communications as an underlay using sequential second price auction," in *Proc., IEEE Intl. Conf. on Communications*, June 2012, pp. 445–449.

[132] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, X. Cheng, and B. Jiao, "Efficiency resource allocation for device-to-device underlay communication systems: a reverse iterative combinatorial auction based approach," *IEEE Journal on Sel. Areas in Communications*, vol. 31, no. 9, pp. 348–358, Sep. 2013.

[133] F. Wang, L. Song, Z. Han, Q. Zhao, and X. Wang, "Joint scheduling and resource allocation for device-to-device underlay communication," in *Proc., IEEE Wireless Networking and Comm. Conf.*, Apr. 2013, pp. 134–139.

[134] T. Chen, G. Charbit, and S. Hakola, "Time hopping for device-to-device communication in LTE cellular system," in *Proc., IEEE Wireless Networking and Comm. Conf.*, Apr. 2010, pp. 1–6.

[135] L. Lei, Y. Zhang, X. Shen, C. Lin, and Z. Zhong, "Performance analysis of device-to-device communications with dynamic interference using Stochastic Petri Nets," *IEEE Trans. on Wireless Communications*, vol. 12, no. 12, pp. 6121–6141, Dec. 2013.

[136] S. Xiang, T. Peng, Z. Liu, and W. Wang, "A distance-dependent mode selection algorithm in heterogeneous D2D and IMT-Advanced network," in *IEEE Globecom Workshops*, Dec. 2012, pp. 416–420.

[137] T. Adachi and M. Nakagawa, "Battery consumption and handoff examination of a cellular ad-hoc united communication system for operational mobile robots," in *Proc., IEEE PIMRC*, vol. 3, Sep. 1998, pp. 1193–1197.

[138] S. Hakola, T. Chen, J. Lehtomaki, and T. Koskela, "Device-to-device (D2D) communication in cellular network-performance analysis of optimum and practical communication mode selection," in *Proc., IEEE Wireless Networking and Comm. Conf.*, Apr. 2010, pp. 1–6.

[139] K. Doppler, C.-H. Yu, C. B. Ribeiro, and P. Janis, "Mode selection for device-to-device communication underlaying an LTE-Advanced network," in *Proc., IEEE Wireless Networking and Comm. Conf.*, Apr. 2010, pp. 1–6.

[140] M. Jung, K. Hwang, and S. Choi, "Joint mode selection and power allocation scheme for power-efficient device-to-device (D2D) communication," in *Proc., IEEE Veh. Technology Conf.*, May 2012, pp. 1–5.

[141] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE Journal on Sel. Areas in Communications*, vol. 27, no. 7, pp. 1029–1046, Aug. 2009.

[142] F. Baccelli, B. Blaszczyszyn, and P. Muhlethaler, "An Aloha protocol for multihop mobile wireless networks," *IEEE Trans. on Info. Theory*, vol. 52, no. 2, pp. 421–436, Feb. 2006.

[143] F. Baccelli, J. Li, T. Richardson, S. Shakkottai, S. Subramanian, and X. Wu, "On optimizing CSMA for wide area ad-hoc networks," in *Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2011, pp. 354–359.

[144] S. P. Weber, X. Yang, J. G. Andrews, and G. De Veciana, "Transmission capacity of wireless ad hoc networks with outage constraints," *IEEE Trans. on Info. Theory*, vol. 51, no. 12, pp. 4091–4102, Dec. 2005.

[145] S. Weber, J. G. Andrews, and N. Jindal, "An overview of the transmission capacity of wireless networks," *IEEE Trans. on Communications*, vol. 58, no. 12, pp. 3593–3604, Dec. 2010.

[146] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. on Communications*, no. 11, pp. 3122 –3134, Nov. 2011.

[147] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE Journal on Sel. Areas in Communications*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[148] Alcatel-Lucent, "LTE D2D dropping and association," *R1-131789*, Apr. 2013.

[149] Qualcomm, "D2D deployment scenarios," *R1-131412*, Apr. 2013.

[150] N. Abramson, "The Aloha system: another alternative for computer communications," in *Proc. Fall Joint Computer Conf. AFIPS,*, Nov. 1970, pp. 281–285.

[151] X. Lin, J. G. Andrews, and A. Ghosh, "Modeling, analysis and design for carrier aggregation in heterogeneous cellular networks," *IEEE Trans. on Communications*, vol. 61, no. 9, pp. 4002–4015, Sep. 2013.

[152] B. Blaszczyszyn, M. K. Karray, and H.-P. Keeler, "Using Poisson processes to model lattice cellular networks," in *Proc., IEEE INFOCOM*, Apr. 2013.

[153] F. Baccelli and B. Blaszczyszyn, *Stochastic geometry and wireless networks*. Now Publishers Inc, 2009, vol. 1.

[154] D. Stoyan, W. S. Kendall, J. Mecke, and L. Ruschendorf, *Stochastic Geometry And Its Applications*. Wiley New York, 1987, vol. 2.

[155] F. Baccelli, C. Gloaguen, and S. Zuyev, "Superposition of planar voronoi tessellations," *Stochastic Models*, vol. 16, no. 1, pp. 69–98, Mar. 2000.

[156] J.-S. Ferenc and Z. Néda, "On the size distribution of Poisson Voronoi cells," *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 518–526, Nov. 2007.

[157] 3GPP, "Technical specification group radio access network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (Release 9)," *TR 36.942, V9.0.1*, Apr. 2010.

[158] ——, "Further advancements for E-UTRA physical layer aspects," *TR 36.814*, Mar. 2010.

[159] ——, "Coordinated multi-point operation for LTE with non-ideal backhaul (Release 12)," *TR 36.874*, Nov. 2013.

[160] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Device-to-device modeling and analysis with a modified Matern Hardcore BS location model," in *Proc., IEEE Globecom*, Dec. 2013, pp. 1825–1830.

[161] Q. Ye, M. Al-Shalash, C. Caramanis, and J. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks," *IEEE Trans. on Communications*, vol. 63, no. 2, pp. 441–454, Feb. 2015.

[162] C.-H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "On the performance of device-to-device underlay communication with simple power control," in *Proc., IEEE Veh. Technology Conf.*, Apr. 2009, pp. 1–5.

[163] J. Gu, S. J. Bae, B.-G. Choi, and M. Y. Chung, "Dynamic power control mechanism for interference coordination of device-to-device communication in cellular networks," in *International Conference on Ubiquitous and Future Networks (ICUFN)*, June 2011, pp. 71–75.

[164] P. Janis, V. Koivunen, C. Ribeiro, J. Korhonen, K. Doppler, and K. Hugl, "Interference-aware resource allocation for device-to-device radio underlaying cellular networks," in *Proc., IEEE Veh. Technology Conf.*, Apr. 2009, pp. 1–5.

[165] L. Cao and H. Zheng, "Distributed spectrum allocation via local bargaining," in *Proc. IEEE SECON*, Sep. 2005, pp. 475–486.

[166] R. Etkin, A. Parekh, and D. Tse, "Spectrum sharing for unlicensed bands," *IEEE Journal on Sel. Areas in Communications*, vol. 25, no. 3, pp. 517–528, Apr. 2007.

[167] M. Razaviyayn, Z.-Q. Luo, P. Tseng, and J.-S. Pang, "A Stackelberg game approach to distributed spectrum management," *Mathematical programming*, vol. 129, no. 2, pp. 197–224, July 2011.

[168] L. Song, D. Niyato, Z. Han, and E. Hossain, "Game-theoretic resource allocation methods for device-to-device communication," *IEEE Wireless*

Communications, vol. 21, no. 3, pp. 136–144, June 2014.

[169] C. Xu, L. Song, Z. Han, D. Li, and B. Jiao, "Resource allocation using a reverse iterative combinatorial auction for device-to-device underlay cellular networks," in *Proc., IEEE Globecom*, Dec. 2012, pp. 4542–4547.

[170] M. J. Osborne, *A Course in Game Theory.* Cambridge, Mass.: MIT Press, 1994.

[171] G. Debreu, "A social equilibrium existence theorem," *Proc. of the National Academy of Sciences*, vol. 38, no. 10, pp. 886–893, Oct. 1952.

[172] K. Fan, "Fixed-point and minimax theorems in locally convex topological linear spaces," *Proc. of the National Academy of Sciences*, vol. 38, no. 2, pp. 121–126, Feb. 1952.

[173] I. L. Glicksberg, "A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points," *Proc. of the American Mathematical Society*, vol. 3, no. 1, pp. 170–174, 1952.

[174] K. W. Shum, K. K. Leung, and C. W. Sung, "Convergence of iterative waterfilling algorithm for Gaussian interference channels." *IEEE Journal on Sel. Areas in Communications*, vol. 25, no. 6, pp. 1091–1100, Aug. 2007.

[175] K. S. Narendra and M. A. Thathachar, *Learning Automata: An Introduction.* Prentice-Hall, 1989.

[176] W. Yu, G. Ginis, and J. M. Cioffi, "Distributed multiuser power control for digital subscriber lines," *IEEE Journal on Sel. Areas in Communications*, vol. 20, no. 5, pp. 1105–1115, June 2002.

[177] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[178] A. Granas, *Fixed Point Theory*. Springer, 2003.

[179] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*. Academic Press, 1992.

[180] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1964, vol. 3.

[181] M. Hong and Z.-Q. Luo, "Distributed linear precoder optimization and base station selection for an uplink heterogeneous network," *IEEE Trans. on Signal Processing*, vol. 61, no. 12, pp. 3214–3228, June 2013.

[182] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proc., IEEE INFOCOM*, Apr. 2014, pp. 1096–1104.

[183] X. Chen and R. Hu, "Joint uplink and downlink optimal mobile association in a wireless heterogeneous network," in *Proc., IEEE Globecom*, Dec. 2012, pp. 4131–4137.

[184] S. Singh, X. Zhang, and J. G. Andrews, "Joint rate and SINR coverage analysis for decoupled uplink-downlink biased cell associations in HetNets," *arXiv preprint arXiv:1412.1898*, Dec. 2014.

[185] Z. Feng, Z. Feng, W. Li, and W. Chen, "Downlink and uplink splitting user association in two-tier heterogeneous cellular networks," in *Proc., IEEE Globecom*, Dec. 2014, pp. 4659–4664.

[186] K. Smiljkovikj, P. Popovski, and L. Gavrilovska, "Analysis of the decoupled access for downlink and uplink in wireless heterogeneous networks," *IEEE Wireless Communications Letters*, vol. PP, no. 99, pp. 1–1, Jan. 2015.

[187] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *Proc., IEEE Globecom*, Dec. 2014, pp. 1798–1803.

[188] K. Balachandran, J. Kang, K. Karakayali, and K. Rege, "Virtual soft handoff enabled dominant interference cancellation for enhanced uplink performance in heterogeneous cellular networks," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Apr. 2012, pp. 109–114.

Vita

Qiaoyang Ye is a Ph.D. candidate in Electrical and Computer Engineering at The University of Texas at Austin. She received the B.Eng. degree in Information Science and Electronic Engineering from Zhejiang University, China, in 2010 and the M.S. in Electrical and Computer Engineering from UT Austin in 2013. Her research interests are in the broad area of wireless communications and networking with current focus on the small cells, massive MIMO, and device-to-device communication. She was an Exemplary Reviewer for IEEE Wireless Communications Letters in 2014. She has held summer internships at Huawei Technologies in Plano, TX in summer 2012 and summer 2013, and DOCOMO Innovations in Palo Alto, CA in summer 2014.

Permanent email: yqy614@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.