# Efficient Sharing of Radio Spectrum for Wireless Networks

**Xu Yuan**

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

Y. Thomas Hou, Chair
Wenjing Lou
Jeffrey H. Reed
Hanif D. Sherali
Yi Shi
Yaling Yang

May 4, 2016
Blacksburg, Virginia

Keywords: Wireless network, spectrum sharing, coexistence, interference, primary network, secondary network, cognitive radio, mathematical modeling, optimization, algorithm design

# Efficient Sharing of Radio Spectrum for Wireless Networks

Xu Yuan

## ABSTRACT

The radio spectrum that can be used for wireless communications is a finite but extremely valuable resource. During the past two decades, with the proliferation of new wireless applications, the use of the radio spectrum has intensified to the point that improved spectrum sharing policies and new mechanisms are needed to enhance its utilization efficiency.

This dissertation studies spectrum sharing and coexistence on both licensed and unlicensed bands for wireless networks. For licensed bands, we study two coexistence paradigms: transparent coexistence (a.k.a., underlay) and policy-based network cooperation (a.k.a., overlay). These two paradigms can offer significant improvement in spectrum utilization and throughput performance than the interweave paradigm. For unlicensed band, we study coexistence of Wi-Fi and LTE, the two most poplar wireless networks. We summarize our contributions as follows.

Under the transparent coexistence paradigm, we study how to achieve optimal spectrum sharing between primary and secondary users in a multi-hop network environment, where the secondary multi-hop wireless network is allowed to use the same spectrum simultaneously with the primary multi-hop wireless network as along as their interference is properly managed. For interference cancelation, we employ MIMO at the secondary nodes. We study the following three problems:

- **Transparent Coexistence: Modeling and Optimization** We develop a rigorous mathematical model to address channel/time slot scheduling, inter-network interference cancelation (IC) between primary and secondary networks, and intra-network IC within the secondary network. As an application, we study a throughput maximization problem with the objective of maximizing the secondary network throughput and show that transparent coexistence paradigm offers significant improvement in spectrum access and throughput performance over the interweave paradigm.

- **Transparent Coexistence: A Distributed Algorithm** Following our efforts on establishing

mathematical model for transparent coexistence, we develop a distributed algorithm to maximize throughput. Our proposed algorithm is based on local information exchange among neighboring nodes. We prove that IC among the nodes by our distributed data structure at each node can be mapped to a global MIMO IC model among all nodes in the network. This is significant as it guarantees the existences of feasible precoding/decoding vectors at all secondary nodes to achieve our desired IC in the network (i.e., feasibility at the PHY layer).

- **Transparent Coexistence: An Online Algorithm** To accommodate dynamic user arrival and departure, we develop an online algorithm for transparent coexistence. Our traffic management algorithm is to address session (flow) level dynamics, i.e., to determine if a new session can be admitted into the network and how to control the additional IC that comes with it. More important, we prove that all inter- and intra-network IC through our DoF allocation is indeed feasible at the PHY layer at all time under traffic dynamics.

Under policy-based network cooperation, the primary and secondary nodes are allowed to cooperate with each other at the node level to relay each other's traffic. We are interesting in exploring how such cooperation can help improve throughput. We study the following two problems:

- **Policy-based Network Cooperation: Modeling and Optimization** We develop a policy-based network cooperation paradigm for efficient spectrum sharing between the primary and secondary users. Such network cooperation can be defined by a set of policies under which different degrees of cooperation are to be achieved. As an example, we study a specific policy called UPS, which allows a complete cooperation between the primary and secondary networks at the node level to relay each other's traffic. Through rigorous mathematical modeling, problem formulation, approximation solution, and simulation results, we show that the UPS offers significantly better throughput performance than that under the interweave paradigm.

- **Policy-based Network Cooperation: The Throughput Region** We are interested in maximizing the achievable throughput for both the primary and secondary networks. We formulate the problem as a multicriteria optimization problem. Through a novel approach based

on weighted Chebyshev norm, we transform the multicriteria optimization problem into a single criteria optimization problem and find a sequence of Pareto-optimal points iteratively. Based on the Pareto-optimal points, we construct the throughput curve and show that it is $\varepsilon$-approximation to the optimal curve. Further, we demonstrate that the throughput region (the area under the throughput curve) under node-level cooperation is substantially larger than that when there is no node-level cooperation.

For spectrum sharing and coexistence on the unlicensed band, we study coexistence of Wi-Fi and LTE, the two most popular wireless networks. We take a novel and neutral approach to understand coexistence between the two technologies from the perspective of user satisfaction. We have two interesting findings: (i) In terms of maximizing total users satisfaction function, there does not appear to be any advantage with coexistence of unlicensed spectrum for Wi-Fi and LTE under static spectrum allocation; (ii) There is significant advantage in coexistence between Wi-Fi and LTE under adaptive spectrum partitioning over Wi-Fi only and static partitioning strategies.

# Acknowledgments

I would like to acknowledge my advisor, members of my dissertation committee, labmates, and my family for their support during my work on this dissertation.

Throughput my Ph.D life, I have benefited greatly from my interactions with many individuals. First and foremost, I want to thank my advisor, Professor Tom Hou, for his guidance, support, and encouragement throughout my Ph.D. studies. Professor Hou spent countless hours on helping me to define my research problems, present my findings, and revise every sentence in my papers. What I learned from him is not just how to solve a problem, but rather the way of conducting research at the highest quality. His keen vision guided me to the outcome of this dissertation, his remarkable efforts contributed to all the important findings in this dissertation, and his pursuit of scholarship inspired me to improve myself to the academic level that I would have never achieved otherwise. His passion to research, his way of thinking, and his scholarly manner inspire me toward a future career in academia.

I want to thank Dr. Yi Shi for his help and guidance. I am indebted to him for many helpful discussions on my research problems and solutions, Whenever I got stuck with a difficult problem, I would always turn to him for rescue. I enjoyed the numerous hours that we spent together solving some hard problems, and I appreciated the lengthy hours that he spent proofreading my papers. These constructive discussions and feedback from him have greatly enhanced the quality of my work in this dissertation.

I would like to thank Prof. Hanif Sherali for taking a genuine interest in my work. His input

has shaped the direction of many solutions in this dissertation. His valuable feedback helped me improve this dissertation in many ways. He is undisputed a role model of research excellence, professionalism, and humanity.

I would like to acknowledge and thank the rest of my dissertation committee for their time and efforts: Professor Wenjing lou, Professor Jeff Reed, and Professor Yaling Yang. Their questions and comments have helped improve the overall quality of my dissertation.

My gratitude extends to my current and former colleagues in the Complex Network and Security Research (CNSR) Lab for their friendship and collaboration with me, including Xiaoqi Qin, Changlai Du, Brian Jalaeian, Amr Nabil, Xiangwei Zheng, Yan Huang, Sushant Sharma, Canming Jiang, Liguang Xie, Huacheng Zeng, Qiben Yan, Rongbo Zhu, Xiaozhu Liu, Lili Zhang, An Li, Nan Jiang, and Feng Tian.

I would like to thank my roommates and friends who helped and assisted me in the past six years. I thank my roommates Peng Lv, Mengna Liu, Yifei Ma, Miao Yao and Jiafei Kuang for sharing a friendly and relaxed iving environment with me. I owe my thanks to my friends Zhenning Cao, Hao Zhang, Junyang Chen, Lei Gao, Jia Guo, Qinghui Mu, Qian Cao, Li Tan and Qing Li. I have thoroughly enjoyed my interactions with them in my life in Blacksburg, and the good social time we had together. They became my extended family during my six years in Blacksburg. Your friendship is greatly appreciated and will be remembered.

Finally, I wish to thank my parents, brother and sisters, who have supported me through the many years of my academic pursuits. They have always been there to inspire and encourage me to the next higher academic degree. No words can adequately express my deepest and utmost love to my parents, brother and sisters, to whom I would like to dedicate this dissertation.

# Funding Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under Grants 1064953, 1443889, 1343222, 1102013, 1247830 and the Office of Naval Research (ONR) under Grant N000141310080. I also acknowledge Virginia Techs Advanced Research Computing for giving me access to the BlueRidge computer cluster.

# Contents

**1 Introduction** **1**

1.1 Background and Motivation . . . . . 1

1.2 Dissertation Outline and Contributions . . . . . 3

**2 Transparent Coexistence: Mathematical Modeling and Optimization** **8**

2.1 Introduction . . . . . 8

2.2 Background and Motivation . . . . . 10

2.2.1 Channel State Information . . . . . 12

2.3 Problem Statement . . . . . 13

2.4 Mathematical Modeling . . . . . 17

2.4.1 Notation . . . . . 17

2.4.2 Node Ordering for IC in Secondary Network . . . . . 19

2.4.3 DoF Allocation at A Secondary Transmitter . . . . . 20

2.4.4 DoF Allocation at A Secondary Receiver . . . . . 23

2.5 Case Study: A Throughput Maximization Problem . . . . . 25

2.5.1 Problem Formulation . . . . . 25

2.5.2 Overview of Solution Algorithm . . . . . 26
2.5.3 Algorithm Details . . . . . 28
2.6 Performance Evaluation . . . . . 30
2.6.1 An Example . . . . . 30
2.6.2 Comparison to Interference Avoidance Paradigm . . . . . 36
2.6.3 Impact of Various System Parameters . . . . . 38
2.6.4 Complete Results . . . . . 42
2.7 Chapter Summary . . . . . 42

**3 Transparent Coexistence: A Distributed Algorithm 43**

3.1 Introduction . . . . . 43
3.2 Problem Statement . . . . . 46
3.3 A Distributed Algorithm . . . . . 49
3.3.1 State Information at Secondary Nodes . . . . . 50
3.3.2 Step 1: Link Selection . . . . . 53
3.3.3 Step 2: Data Stream Increment . . . . . 54
3.3.4 Step 3: Adjusting Node's IC Responsibility . . . . . 58
3.4 Physical Layer Feasibility . . . . . 63
3.5 Complexity Analysis . . . . . 70
3.6 Simulation Results . . . . . 71
3.6.1 Simulation Setting . . . . . 71
3.6.2 A Case Study . . . . . 72

3.6.3 Comparison to Interweave Paradigm . . . . . 76
3.6.4 Complete Results . . . . . 77
3.7 Chapter Summary . . . . . 79

**4 Transparent Coexistence: An Online Algorithm 80**

4.1 Introduction . . . . . 80
4.2 Transparent Coexistence: A Primer . . . . . 83
4.3 Problem Statement . . . . . 87
4.4 An Online Algorithm . . . . . 89
4.4.1 Data Structure at Secondary Node . . . . . 90
4.4.2 Initiation of a New Secondary Session . . . . . 93
4.4.3 Termination of a Secondary Session . . . . . 95
4.4.4 Initiation of a New Primary Session . . . . . 97
4.4.5 Termination of a Primary Session . . . . . 99
4.4.6 Coping the Race Problem . . . . . 100
4.5 Physical Layer Feasibility . . . . . 102
4.6 Performance Evaluation . . . . . 105
4.6.1 Parameter Settings . . . . . 105
4.6.2 Lost Secondary Sessions . . . . . 105
4.6.3 Validation of Transparent Coexistence . . . . . 112
4.7 Chapter Summary . . . . . 118

**5 Policy-based Network Cooperation: Mathematical Modeling and Optimization 119**

5.1 Introduction . . . 119
5.2 Related Work . . . 122
5.3 The Case of Policy-based Network Cooperation . . . 123
5.4 Case Study: UPS Policy . . . 126
5.4.1 Problem Statement . . . 126
5.4.2 Mathematical Modeling . . . 127
5.4.3 Problem Formulation . . . 132
5.5 An Approximate Solution . . . 133
5.5.1 Overview . . . 133
5.5.2 Linearization . . . 133
5.5.3 Approximation Gap . . . 136
5.6 Simulation Results . . . 138
5.6.1 Simulation Setting . . . 139
5.6.2 An Example . . . 139
5.6.3 Varying the Number of Nodes . . . 148
5.6.4 Varying Session Numbers . . . 149
5.7 Chapter Summary . . . 153

**6 Policy-based Network Cooperation: Throughput Region 154**

6.1 Introduction . . . 154
6.2 Mathematical Modeling and Formulation . . . 158
6.2.1 Network Model . . . 158

6.2.2 Interference Modeling . . . 160
6.2.3 Traffic Modeling . . . 161
6.2.4 Multiobjective Formulation . . . 163
6.3 An Approximation Algorithm . . . 164
6.3.1 Background and Roadmap . . . 164
6.3.2 Single Objective Formulation with Chebyshev Norm . . . 166
6.3.3 Finding Pareto-optimal Point for a Given $\beta$ . . . 168
6.3.4 Determination of New Pareto-optimal Points . . . 169
6.3.5 Main Result . . . 179
6.4 A Case Study . . . 181
6.4.1 Simulation Setting . . . 181
6.4.2 Throughput Curve . . . 182
6.4.3 Comparison to Other Paradigms . . . 187
6.5 Chapter Summary . . . 188

**7 Coexistence between Wi-Fi and LTE on Unlicensed Spectrum 189**

7.1 Introduction . . . 189
7.2 Network Architecture . . . 192
7.3 Scenario A: Wi-Fi Only . . . 195
7.4 Scenario B: Coexistence Through Static Spectrum Partitioning . . . 197
7.4.1 Mathematical Modeling . . . 197
7.4.2 Problem Formulation . . . 199

7.4.3 Reformulation . . . . . 200
7.5 Scenario C: Coexistence Through Adaptive Spectrum Partitioning . . . . . 203
7.6 Performance Evaluation . . . . . 205
7.6.1 Parameter Setting . . . . . 205
7.6.2 Comparison Under Different Satisfaction Coefficients . . . . . 207
7.6.3 Different Bandwidth Allocation in Static Partitioning Scheme . . . . . 210
7.6.4 Varying Traffic Load . . . . . 210
7.7 Semi-Adaptive Algorithm for Practical Implementing . . . . . 213
7.7.1 Motivation . . . . . 213
7.7.2 Algorithm Design . . . . . 213
7.7.3 Performance Evaluation . . . . . 216
7.8 Related Work . . . . . 219
7.9 Chapter Summary . . . . . 220

**8 Dissertation Summary and Future Work** **222**

8.1 Dissertation Summary . . . . . 222
8.2 Future Work . . . . . 224

**Bibliography** **226**

# List of Figures


1.1 Interference avoidance in time, frequency, or space domain [3]. . . . . 2

1.2 The structure of this dissertation. . . . . 4

2.1 A simple example illustrating the benefits of using MIMO to allow simultaneous activation of primary and secondary nodes. . . . . 11

2.2 CSI estimation at secondary node $S_1$. . . . . 14

2.3 A multi-hop secondary network co-located with a multi-hop primary network. . . . 15

2.4 An example illustrating how to fix $x$, $y$, and some $z$ variables in Phase II. . . . . 30

2.5 Active sessions in the primary and secondary networks. . . . . 33

2.6 Channel and time slot scheduling on each link for the secondary sessions by our solution algorithm. Channel and time slot scheduling on each link for the primary sessions are given in Fig. 2.5. . . . . 33

2.7 Illustration of interference relationships among the primary and secondary links on channel 1 in time slot 2 in the case study. . . . . 37

2.8 Channel and time slot scheduling on each link for the secondary sessions under the interference avoidance paradigm. . . . . 38

2.9 Impact of the various system parameters on the performance of transparent coexistence and interference avoidance paradigms. . . . . 39

3.1 A multi-hop secondary network co-located in the same area as a multi-hop primary network. . . . . 46

3.2 A simple example illustrating SM and IC. A solid line represents the primary link, a dashed line represents a secondary link, and a dotted line represents an interference. 47

3.3 Maintaining two local sets at node $i$ to distinguish IC responsibility between node $i$ and its neighboring nodes. . . . . 52

3.4 Four cases of link status. . . . . 54

3.5 Pseudocode to update state information when $s_i(t) =$ Idle. . . . . 56

3.6 Pseudocode to update state information when $s_j(t) =$ Idle. . . . . 57

3.7 Pseudocode to update state information when $s_i(t) =$ Tx. . . . . 57

3.8 Pseudocode to update state information when $s_j(t) =$ Rx. . . . . 58

3.9 Determining the eligibility of receive node $a \in \mathcal{B}_k(t)$ when node $k$ is a transmit node. . . . . 60

3.10 An illustration of movement process when node $k$ is an idle node. . . . . 62

3.11 Update state information at $k$. . . . . 64

3.12 Update state information at $a$. . . . . 64

3.13 A secondary transmit node $i$ performs IC to neighboring primary and secondary receive nodes in a time slot $t$. . . . . 68

3.14 Routing topology for each primary and secondary sessions and scheduling on each link of the respective route. The numbers in the box next to a link show the time slots when the link is active. . . . . 73

3.15 Active links in time slot $6$ in both primary and secondary networks. . . . . 73

3.16 A global node ordering for IC in time slot $6$. . . . . 76

3.17 Routing for each session and scheduling on each link for both primary and secondary networks under the interweave paradigm. . . . . 77

4.1 The underlay coexistence of one secondary link with one primary link. A solid line represents a primary link, a dashed line represents a secondary link, and a dotted line represents a interference. . . . . 84

4.2 The multi-hop primary and secondary networks. . . . . 87

4.3 The locations of the primary and secondary nodes. . . . . 106

4.4 Cumulative secondary arrivals, admitted secondary arrivals by offline algorithm, and admitted secondary arrivals by our online algorithm. Both primary and secondary session arrival rates are 1 per minute. . . . . 107

4.5 The number of active primary sessions in the network, the number of secondary sessions that can be admitted into the network by the offline algorithm, and the number of secondary sessions that are admitted into the network by our online algorithm, all over a 2-hour period. . . . . 109

4.6 Cumulative secondary arrivals, admitted secondary arrivals by offline algorithm, and admitted secondary arrivals by our online algorithm. Primary and secondary session arrival rates are 1 and 5, respectively. . . . . 110

4.7 Cumulative secondary arrivals, admitted secondary arrivals by offline algorithm, and admitted secondary arrivals by our online algorithm. Primary and secondary session arrival rates are 1 and 10, respectively. . . . . 110

4.8 Ratios between admitted secondary sessions by our online algorithm and that by the offline algorithm with different secondary sessions arrival rate. . . . . 111

4.9 The CDFs of computation time by the offline algorithm when the secondary sessions arrival rates are 1, 5, and 10 respectively. The cutoff termination time for the offline algorithm is set to 1 hour. . . . . 111

4.10 The scheduling and routing before and after the new secondary session $S_{42} \rightarrow S_{10}$ arrives. . . . . 114

4.11 Interference relationship in the first two time slots before new secondary session $S_{42} \rightarrow S_{10}$ arrives. . . . . 115

4.12 Interference relationship in each time slot after new secondary session $S_{42} \rightarrow S_{10}$ arrives. . . . . 116

5.1 Network topologies under the interweave and the UPS policy. . . . . 123

5.2 Piece-wise approximation with line segments. . . . . 134

5.3 An illustration of the maximum approximation error for piece-wise line segment. . 135

5.4 A Region 1 example that showing the flow routing topologies and scheduling for the primary and secondary sessions, where the solid line segments are for the primary sessions while the dashed line segments are for the secondary sessions. . . . 145

5.5 A Region 2 example that showing the flow routing topologies and scheduling for the primary and secondary sessions. . . . . 146

5.6 A Region 3 example that showing the flow routing topologies and scheduling for the primary and secondary sessions in the UPS policy. . . . . 147

5.7 The locations of the source and destination nodes of the primary and secondary sessions. . . . . 148

5.8 The comparison of the SS utility objectives for different number of nodes ($K = 10, 15, 20, 25,$ and $30$) with the increasing rate requirements for the primary sessions. 150

5.9 A 20-node primary network and a 20-node secondary network. . . . . 150

6.1 An illustration of the UPS policy for multi-hop primary and secondary networks. . 156

6.2 Assuming $K$ does not fall between $A$ and $B$. . . . . 172

6.3 Assuming $K$ does not fall between $A$ and $B$. . . . . 178

6.4 Pseudo-code of an approximation algorithm to find $(1-\varepsilon)$-optimal throughput curve. 180

6.5 The Pareto-optimal point $R$ is represented by $D_1$ (or $D_2$) with $\varepsilon$-approximation. . . 181

6.6 The locations of a 15-node primary network and a 15-node secondary network. . . 182

6.7 The throughput curve found by our algorithm. . . . . 185

6.8 A comparison of the throughput region under the UPS policy and the interweave paradigm. . . . . 187

7.1 The coexistence of Wi-Fi and LTE in a picocell-sized area. . . . . 191

7.2 A cloud-based control plane that coordinates spectrum sharing between Wi-Fi APs and LTE BS. . . . . 193

7.3 One LTE BS and multiple Wi-Fi APs that are randomly deployed in a circle with radius 100. . . . . 206

7.4 Maximum users satisfaction under Wi-Fi only, static spectrum partitioning, and adaptive spectrum partitioning with different satisfaction coefficients. . . . . 208

7.5 Normalized users satisfaction of Wi-Fi only and static spectrum partitioning with respect to adaptive spectrum partitioning. . . . . 209

7.6 Normalized users satisfaction of Wi-Fi only and static spectrum partitioning under different bandwidth allocation with respect to those for adaptive spectrum partitioning. . . . . 211

7.7 Normalized users satisfaction of Wi-Fi only and static spectrum partitioning with respect to those of adaptive spectrum partitioning when the user arrival rates are 10, 30, and 50 per hour. . . . . 212

7.8 Normalized objective value for the proposed semi-adaptive algorithm to fully adaptive spectrum partitioning with different user arrival rates. . . . . 217

7.9 The CDFs of normalized objective values for the proposed semi-adaptive algorithm to fully adaptive spectrum partitioning under different user arrival rates. . . . . 218

# List of Tables


2.1 Notation . . . . . 18

2.2 Location of each node for the 20-node primary network and 30-node secondary network. . . . . 31

2.3 Source and destination nodes of each session in the primary and secondary networks. 32

2.4 Channel and time slot scheduling on each link, DoF allocation for SM, and throughput on each link for the secondary sessions. . . . . 34

2.5 DOF allocation for SM and IC on $(b, t) = (1, 2)$ at each node in the secondary network. . . . . 35

2.6 Channel and time slot scheduling on each link, DoF allocation for SM, and link rate on each link for the secondary sessions under the interference avoidance paradigm. 40

2.7 Achievable minimum session throughput under transparent coexistence paradigm and interference avoidance paradigm for 50 cases. . . . . 41

3.1 State information at each node $i$ . . . . . 51

3.2 DoF allocation for SM and IC at each active secondary node in time slot $6$. . . . . 75

3.3 Results for 50 network instances. . . . . 78

4.1 DoF allocation for SM and IC for the secondary sessions in each time slot before the new session arrives. . . . 113

4.2 DoF allocation for SM and IC for the secondary sessions in each time slot after the new session arrives. . . . 117

5.1 Notation for UPS paradigm . . . 128

5.2 Location of primary and secondary nodes for the 30-node network. . . . 140

5.3 The source and destination nodes for each session in the 30-node network. . . . 140

5.4 The approximation gap between the SS utility objectives of linearized problem and original problem. . . . 142

5.5 Performance comparison between the UPS policy and the interweave paradigms for different primary session rate requirements. . . . 143

5.6 The average SS utility objectives for different $K$ users. . . . 149

5.7 Feasibility performance of the primary sessions and utilities of the secondary sessions under increasing number of the primary sessions. . . . 151

5.8 Secondary sessions' utility values under increasing number of the secondary sessions. 152

6.1 Notation . . . 159

6.2 New Pareto-optimal point that is found by two known Pareto-optimal points in each iteration. "PO" represents Pareto-optimal points. . . . 183

7.1 Notation . . . 194

7.2 The constants and optimization variables in the formulation of Wi-Fi only, static spectrum partition, and adaptive spectrum partition. . . . 205

# Chapter 1

# Introduction

## 1.1 Background and Motivation

The last decade has witnessed rapid advance in research and development of spectrum sharing technologies. Recent report by the President's Council of Advisors on Science and Technology (PCAST) [46] called for the sharing of 1 GHz of federal government radio spectrum with non-government entities to spur economic growth. This report further accelerated the pace of commercialization of innovative spectrum sharing technologies. A number of grand challenges has been raised, which include: (1) accessing the economic trade-offs of incentivizing spectrum sharing under multiple scenarios; (2) devising models and process that can operate on huge datasets of wireless feedback, rapidly access spectrum usage, and adjust spectrum sharing parameters in real-time; (3) understanding what data from spectrum usage can be collected and analyzed to access spectrum utilization without infringing on users' privacy; (4) facilitating radio-frequency propagation measurements that provide a baseline for expected sharing efficiency in different bands; (5) researching protocols, policies, models, and frameworks to enable future spectrum sharing architectures. In this dissertation, we focus on (5) and specifically on researching new policies, models and frameworks to enable future spectrum sharing architectures.

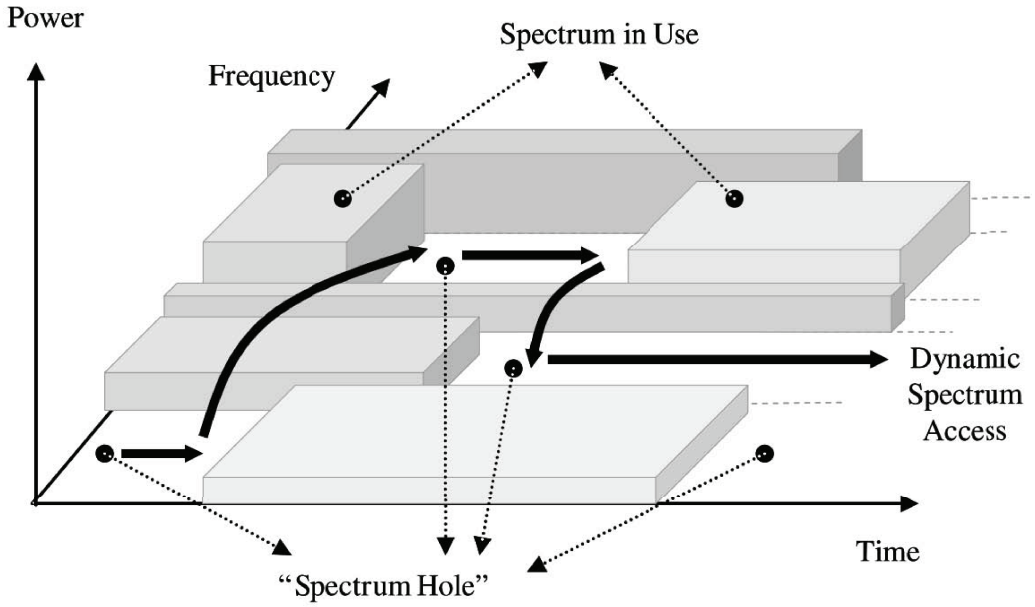



Figure 1.1: Interference avoidance in time, frequency, or space domain [3].

On the licensed spectrum, there has been extensive research on exploring spectrum sharing and coexistence between primary and secondary networks in recent years. In [22], Goldsmith *et al.* outlined three spectrum sharing paradigms for cognitive radios (CR), namely underlay, overlay, and interweave. These three paradigms were defined from an information theoretic perspective, solely based on how much side information (e.g., channel conditions, codebooks) is available to the CRs. In the networking community, these three paradigms have been mapped into specific scenarios of how primary and secondary networks interact with each other for data forwarding. Specifically, the interweave paradigm follows traditional interference avoidance, which refers to that the secondary nodes are allowed to use a spectrum allocated to the primary nodes only when the primary nodes do not use it in the same time, frequency, or space (see Figure 1.1) [21,26,72]. This is in analogy to the classic interference avoidance in medium access, or in CR terminology, dynamic spectrum access (DSA). This is the prevailing paradigm on which most of the research efforts have been devoted by the CR community in recent years. The underlay paradigm refers to that secondary users' activities or interference on primary users is negligible (or below a given threshold). In contrast to the interweave paradigm, secondary users may be active concurrently with the primary users

in the same vicinity and in the same frequency. Potential interference from secondary users may be properly canceled (by secondary users) via various interference cancellation (IC) techniques so that residual interfering signals by secondary users are negligible to the primary users [23, 33, 85, 86]). Finally, the overlay paradigm requires that secondary users have primary users codebook and messages so that secondary users can help maintain or improve the communication of primary users while still achieving some communication on their own [31,42,61,79,83]. From a networking perspective, the overlay paradigm can be interpreted as having secondary users help forward traffic of primary users on top of its own communications.

On the unlicensed bands, there has been great interest from the cellular service providers to use unlicensed spectrum for their service offerings. However, existing unlicensed users in these bands (e.g., Wi-Fi in 5 GHz band) have serious concern that such coexistence will jeopardize their service quality. In [29,53,69], experimental results showed that Wi-Fi throughput may be reduced by 90% when interfered by LTE. This is unfair to Wi-Fi and has led to protest by the Wi-Fi Alliance. An efficient and fair coexistence approach between Wi-Fi and LTE remains to be found.

The goal of this dissertation is to address different spectrum sharing approaches for both licensed and unlicensed bands. For the licensed bands, we focus on underlay and overlay paradigms for the coexistence of multi-hop primary and secondary networks. For unlicensed bands, we focus on interweave coexsitence between Wi-Fi and LTE.

## 1.2 Dissertation Outline and Contributions

This dissertation proposes the new spectrum sharing policies and coexistence approaches for enhancing spectrum utilization on both licensed and unlicensed bands. An illustration of the structure of this dissertation is given in Fig. 1.2. The main contributions of each chapter are summarized as follows:

- In Chapter 2, we explore the "transparent coexistence" for spectrum sharing between primary and secondary nodes in a multi-hop network environment. Under this paradigm, the

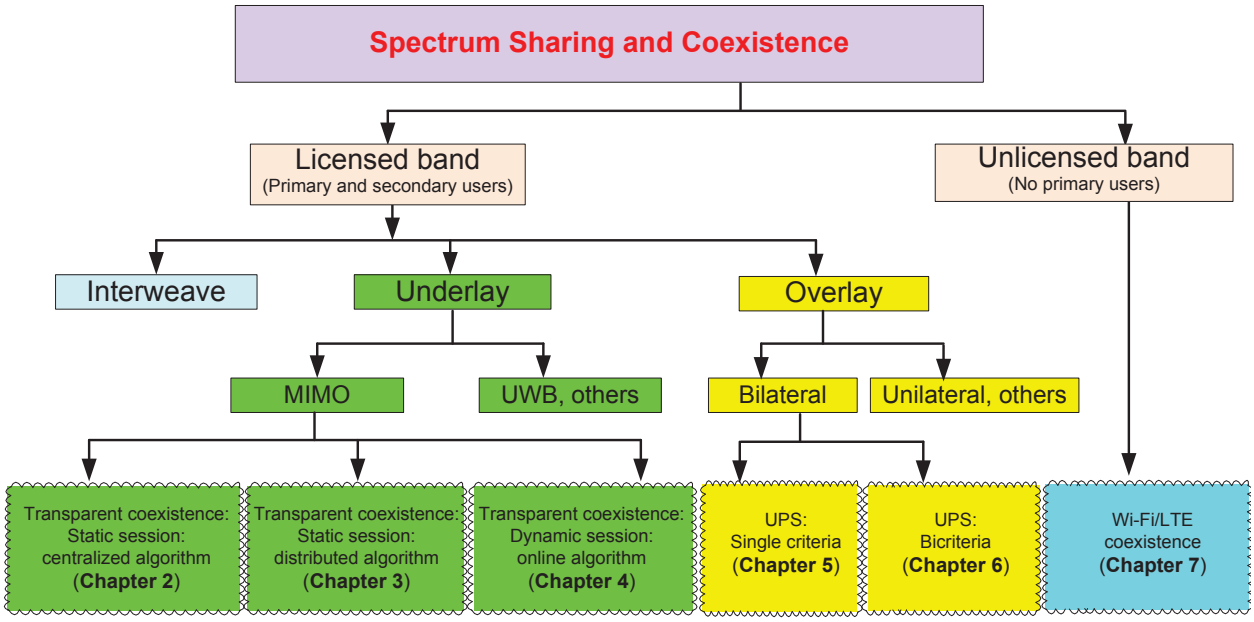



Figure 1.2: The structure of this dissertation.

secondary network is allowed to use the same spectrum simultaneously with the primary network as long as their activities are "transparent" (or "invisible") to the primary network. Such transparency is accomplished through a systematic interference cancelation (IC) by the secondary nodes without any impact on the primary network. Although such a paradigm has been studied in the information theory (IT) and communications (COMM) communities, it is not well understood in the wireless networking community, particularly for multi-hop networks. The new technical challenges in a multi-hop network include channel/time slot scheduling, inter-network IC between primary and secondary networks, and intra-network IC within the secondary network. We develop a rigorous mathematical model for a secondary multi-hop network in the transparent coexistence paradigm. As an application, we apply this model to study a throughput maximization problem and develop an efficient polynomial time algorithm. We offer simulation results to show the significant improvement in spectrum access and throughput performance when compared to interference avoidance paradigm.

- In Chapter 3, we design a distributed iterative algorithm to achieve transparent coexistence

for multi-hop primary and secondary networks. The main challenge in this algorithm is to ensure that IC is done efficiently (i.e., canceled once by a secondary node) and in a feasible manner (i.e., implementable at the PHY layer). In contrary to a centralized IC algorithm which relies on a global node ordering for IC, we only maintain two local sets for each node to keep track of the node's IC responsibilities. We show how to establish, maintain, and update these two local sets for each node in our distributed algorithm. Our distributed algorithm increases the data stream on each active link iteratively based on local computation. Since the nodes in the two local set of a node directly affect the node's IC responsibility, our algorithm attempts to switch nodes in the two sets if it can improve the IC structure. Although no explicit node ordering is maintained in our distributed algorithm, we prove that our distributed data structure at each node (with the use of two local sets) can be mapped to an explicit global node ordering for IC among all nodes in the network. This guarantees the existences of feasible precoding/decoding vectors at the secondary nodes to achieve our desired IC in the network (i.e., feasibility at the PHY layer). Our algorithm is iterative in nature and all steps can be accomplished based on local information exchange among the neighboring nodes. We present simulation results to show that the performance of our distributed algorithm is highly competitive when compared to an upper bound solution from the corresponding centralized problem.

- In Chapter 4, we design an online distributed algorithm to handle dynamic session arrival and departure in the transparent coexistence paradigm. For IC, we again employ multiple antennas on the secondary nodes. Since it takes time to configure the precoding/decoding vectors at a secondary node for spatial multiplexing (SM) and IC, per packet level dynamic traffic management does not appear to be practical. Instead, our traffic management algorithm is to address session (flow) level dynamics, i.e., to determine if a new session can be admitted into the network and how to perform the additional IC that comes with it. Our algorithm ensures that all interferences to/from the primary network and interference within the secondary network are canceled properly so that data transport is free of interference in both the primary and secondary networks at any time. More importantly, we prove that such

inter-network and intra-network IC through our DoF allocation is indeed feasible at the PHY layer at all time under traffic dynamics. We conduct extensive performance evaluation under various traffic loads to show that our online algorithm offers competitive performance when compared to an offline centralized algorithm.

- In Chapter 5, we studies a new and bold spectrum-sharing paradigm beyond the state of the art for future wireless networks. We explore network cooperation as a new dimension for spectrum sharing between the primary and secondary users. Such network cooperation can be defined as a set of policies under which different degrees of cooperation are to be achieved. There are many possible node-level cooperation policies that one can employ under this paradigm. For the purpose of performance study, we consider a specific policy called United cooperation of Primary and Secondary networks (UPS), which allows a complete cooperation between the primary and secondary networks at the node level to relay each other's traffic. We study a problem with the goal of supporting the rate requirement of the primary network traffic while maximizing the throughput of the secondary sessions. For this problem, we develop an optimization model and formulate a combinatorial optimization problem. We also develop an approximation solution based on a piece-wise linearization technique. Through simulation results, we show that the UPS offers significantly better throughput performance than that under the interweave paradigm.

- In Chapter 6, we have an in-depth study of the UPS paradigm in terms of its optimal throughput curve - the maximum achievable throughput for both primary and secondary users. We formulate the problem as a multicriteria optimization problem with the goal of maximizing the throughput of both the primary and secondary users. Through a novel approach based on a weighted Chebyshev norm, we transform the multicriteria optimization problem into a single criteria optimization problem and iteratively find a sequence of Pareto-optimal points. We show that the throughput curve (by connecting consecutive known Pareto-optimal points via "L"-shaped line segments) is $(1-\varepsilon)$-optimal. Through a case study, we show that the throughput region for the UPS paradigm can be substantially larger than that in the interweave paradigm. In addition to demonstrating the large throughput region of the UPS

paradigm, the throughput curve offers a complete landscape of achievable throughput for the primary and secondary users.

- In Chapter 7, we study the coexistence of Wi-Fi and LTE on the unlicensed band. Although there are some proposals on how to achieve coexistence, there remain issues and skepticism. Instead of taking a side in this debate, we take a novel and neutral approach to understand coexistence between Wi-Fi and LTE from the perspective of user satisfaction. Through mathematical modeling, problem formulation and extensive simulations studies, we show that in terms of maximizing total users satisfaction function, there does not appear to be any advantage with coexistence of unlicensed spectrum for Wi-Fi and LTE under static partition of unlicensed spectrum. This finding serves as a powerful counter argument to some telecom carriers' proposal to use the unlicensed spectrum through static partitioning of the unlicensed band for Wi-Fi and LTE. On the other hand, there is a significant advantage in coexistence between Wi-Fi and LTE under adaptive spectrum allocation. Since adaptive spectrum allocation may require a user to change its service provider whenever there is a change among the users, we propose a practical (semi-adaptive) algorithm to implement adaptive spectrum allocation without affecting existing users' service providers. Through performance evaluation, we show that the proposed semi-adaptive algorithm is highly competitive when compared to adaptive spectrum allocation.

# Chapter 2

# Transparent Coexistence: Mathematical Modeling and Optimization

## 2.1 Introduction

Recent push by the government agencies to share federal government radio spectrum with non-government entities has fueled the development of innovative technologies for spectrum sharing [46]. The current prevailing spectrum-sharing paradigm is that secondary nodes (typically equipped with cognitive radios (CRs)) are allowed to use a spectrum channel allocated to the primary nodes only when such a use will not cause interference to the primary nodes [3, 21, 26, 58]. This is also called "interweave" paradigm in [22], which we call *interference avoidance* paradigm in this chapter. Under this paradigm, the wireless networking community has invested significant research efforts in algorithm design and protocol implementation to optimize secondary CR users' performance while ensuring that their activities will not interfere with the primary users.

On the other hand, in the information theory (IT) community, there is a strong interest in exploring information theoretic limit of CR [22]. In particular, researchers have been exploring the potential of *simultaneous activation* of a secondary network with the primary network, as long

as the interference produced by secondary nodes can be properly "controlled" (e.g., canceled) by the secondary nodes. Here, secondary nodes are allowed to access the spectrum as long as they can cancel their interference to the primary nodes in such a way that the primary nodes do not feel the presence of the secondary nodes. In other words, activities by the secondary nodes are made transparent (or "invisible") to the primary nodes. We call this *transparent coexistence* paradigm (also called "underlay" paradigm in [22]) in this chapter. Under this paradigm, secondary nodes are assumed to have powerful (physical layer) capabilities to perform interference cancelation (IC), thereby, allowing them to access the spectrum in a much more aggressive manner than the interference avoidance paradigm.

Although the idea of the transparent coexistence paradigm has been explored in the IT community, results from the IT and communications (COMM) communities have mainly limited to very simple network settings, e.g., several nodes or link pairs, all for *single-hop* communications [5, 23, 33, 85, 86]. The more difficult problem of how transparent coexistence can be achieved in a *multi-hop* secondary network remains open. As shown in [26, 58], the problem complexity associated with multi-hop CR networks is much higher than single-hop CR networks. To date, there are no prior results on transparent coexistence for a multi-hop CR networks.

The goal of this chapter is to advance the theoretical foundation of transparent coexistence paradigm for a multi-hop secondary CR network. We study how a multi-hop secondary CR network can co-exist with a primary network transparently. For IC, we assume that each secondary node is equipped with multiple transmit/receive antennas (MIMO).[1] For a set of channels owned by the primary networks, the primary nodes may use them in whatever manner to suit their needs. On the other hand, the secondary nodes are only allowed to use these channels if they can cancel their interference to the primary nodes. Further, to ensure successful transmission among the secondary nodes, the secondary nodes also need to perform IC to/from the primary nodes as well as potential interference among the secondary nodes. Simply put, all IC burden should rest solely on the secondary nodes and remain invisible to the primary nodes. For this paradigm, we offer a mathematical modeling of channel/time slot scheduling, IC between primary and secondary nodes,

[1] Other IC techniques may also be employed and will be explored in our future studies.

and IC within the secondary network. Based on this model, we study a throughput maximization problem (with the objective of maximizing the minimum throughput among all sessions in the secondary network) without any impact on the primary users. Since the problem has a mixed-integer linear program (MILP) formulation, we develop an efficient solution based on a *sequential fixing* (SF) technique. Through simulation results, we demonstrate how the transparent coexistence paradigm can offer much improved spectrum access and throughput performance than the current interference avoidance paradigm.

The remainder of this chapter is organized as follows. In Section 2.2, we give essential background on how IC may be performed by MIMO. Section 2.3 describes our problem and key challenges. In Section 6.2, we present a mathematical model for the transparent coexistence paradigm where both the primary and secondary networks are multi-hop. Based on this model, in Section 2.5, we study a throughput maximization problem and presents an efficient solution algorithm. Section 6.4 presents simulation results and demonstrates the significant improvement in spectrum access and throughput performance under the transparent coexistence paradigm. Section 2.7 concludes this chapter and discusses the further work.

## 2.2 Background and Motivation

We give a brief review of MIMO in terms of its spatial multiplexing (SM) and IC capabilities [13, 30, 65, 68]. Other capabilities such as spatial diversity [90] are not explored in this chapter and will be considered in our future work.

A simple representation of MIMO can be built upon the so-called degree-of-freedom (DoF) concept [30, 68]. Simply put, the total number of DoFs at a node (no more than the number of antenna elements) represents the available resources at the node. A DoF can be used for either data transmission/reception or IC. Typically, transmitting one data stream requires one DoF at a transmitter and one DoF at its receiver. SM refers to the scenario where multiple DoFs are used to transmit multiple data streams, thus substantially increasing data throughput between the two

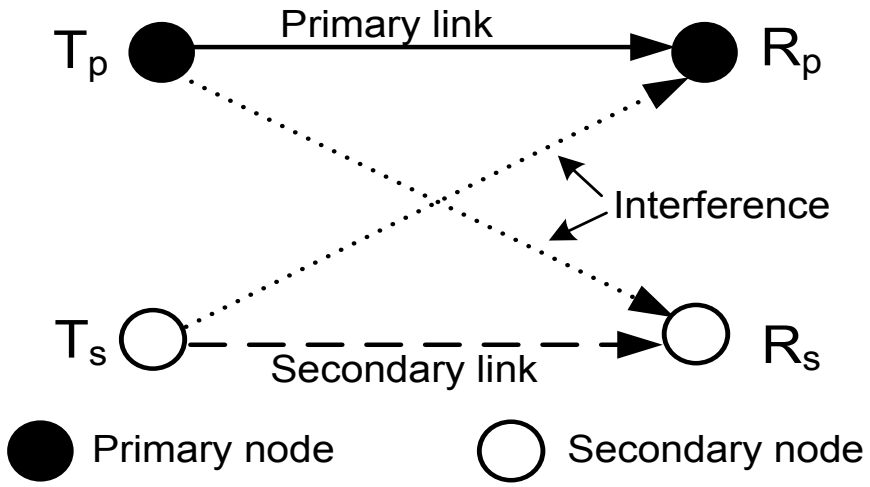



Figure 2.1: A simple example illustrating the benefits of using MIMO to allow simultaneous activation of primary and secondary nodes.

nodes. On the other hand, IC refers to a node's capability to use some of its DoFs to cancel interference, either as a transmitter or as a receiver. Depending on whether IC is done at a transmitter or receiver, the number of required DoF consumption may be different.

- **IC by Tx.** If a transmitter (Tx) is to cancel its interference to an unintended receiver, the number of DoFs required at this transmitter is equal to the number of data streams (or DoFs) that the unintended receiver is trying to receive from its transmitter.
- **IC by Rx.** If a receiver (Rx) is to cancel the interference from an interfering transmitter, the number of DoFs required at this receiver is equal to the number of data streams (or DoFs) that the interfering transmitter is trying to transmit to its intended receiver.

At any node, the sum of DoFs used for SM and IC cannot exceed the total number of DoFs at the node.

A MIMO node's ability to use a subset of its DoFs to cancel interference while to use the remaining subset of DoFs for data transmission allows the possibility of simultaneous activation of the secondary nodes with the primary nodes. We use a simple example to illustrate this point. In Fig. 2.1, suppose $T_p$ and $R_p$ are a pair of transmit and receive nodes in the primary network, while

$T_s$ and $R_s$ are a pair of transmit and receive nodes in the secondary network. Assume that all nodes share the same channel. Suppose $T_p$ is transmitting 1 data stream to $R_p$. Under the interference avoidance paradigm, secondary transmit node $T_s$ is prohibited from transmission on the same channel as it will interfere with primary receive node $R_p$. However, when MIMO is employed on the secondary nodes, simultaneous transmissions can be achieved. Assume secondary nodes $T_s$ and $R_s$ are each equipped with 4 antennas (4 DoFs). $T_s$ can use 1 of its DoFs to cancel its interference to $R_p$ so that $R_p$ can receive its 1 data stream correctly from $T_p$. At node $R_s$, $R_s$ can use 1 of its DoFs to cancel interference from $T_p$. After IC, both $T_s$ and $R_s$ still have 3 DoFs remaining, which can be used for SM of 3 data stream from $T_s$ to $R_s$.

### 2.2.1 Channel State Information

As the above example shows, under transparent coexistence, all IC burden rests upon the secondary nodes. Specifically, a secondary transmit node needs to cancel its interference to all neighboring primary receive nodes who are interfered by this secondary transmitter; a secondary receive node needs to cancel interference from all neighboring primary transmit nodes that interfere with this secondary receiver. To achieve transparency to the primary nodes, it is important for the secondary nodes to have accurate channel state information (CSI). The problem is: how can a secondary node obtain the CSI between itself and its neighboring primary nodes while remaining transparent to the primary nodes?

We propose the following solution to resolve this problem. For each primary node, it typically sends out a pilot sequence (training sequence) to its neighboring primary nodes so that those primary nodes can estimate the CSI. This is the practice for current cellular networks and we assume such a mechanism is available for a primary network. Since we consider a multi-hop network, where each node will act as a transmitter in one time slot but as a receiver in another time slot. Then, each secondary node can *overhear* the pilot sequence signal from the primary node while staying transparent. For example, in Fig. 2.2(a), in time slot $t_1$, when $P_1$ is transmitting the pilot sequence, a secondary node $S_1$ can overhear this sequence from $P_1$. Likewise, in Fig. 2.2(b), in

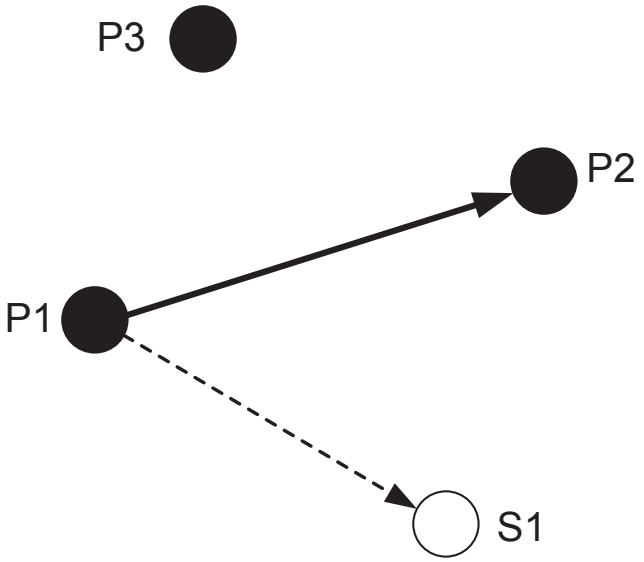
time slot $t_2$, when $P_2$ is transmitting its pilot sequence, the secondary node $S_1$ can overhear this pilot sequence from $P_2$. Suppose the pilot sequence from the primary nodes is publicly available (as in cellular networks) and is known to the secondary nodes. Then the secondary node $S_1$ can use this information and the actual received pilot sequence signal from the primary nodes for channel estimation. Based on the reciprocity property of a wireless channel [63], a secondary node $S_1$ will be able to estimate the CSI in both directions to/from $P_1$ and $P_2$. Likewise, the CSI among the secondary nodes may be derived following a similar approach.

## 2.3 Problem Statement

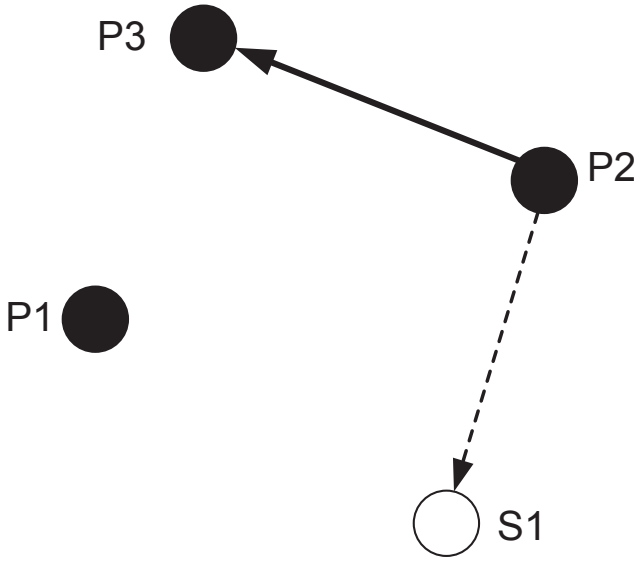
We consider a primary multi-hop ad hoc network $\mathcal{P}$ shown in Fig. 2.3, which is co-located with a secondary multi-hop network $\mathcal{S}$ in the same geographical region. Suppose that there is a set of channels $\mathcal{B}$ owned by the primary network. For scheduling on each channel, we consider a time frame with $T$ equal-length time slots. The primary nodes can use this set of channels and time slots freely as if they were the only nodes in the network. The primary nodes are assumed to be single-antenna nodes. For the secondary nodes, they are allowed to use a time slot $t$ ($1 \leq t \leq T$) on a channel only if their interference to the primary nodes are canceled properly, with complete transparency to the primary nodes. For IC, we assume that the secondary nodes are equipped with MIMO. Some key assumptions that we make in this chapter are the following:

- In primary network, we assume that each primary node is a single-antenna node.[2]
- The secondary nodes need to know the primary nodes' transmission behavior (link scheduling). We assume this information can be derived by the secondary nodes through monitoring/sensing of the primary nodes' activities.
- The secondary nodes need to have CSI to perform IC (to/from the primary nodes and within the secondary nodes). A proposed solution was given in Section 2.2.1.

[2]The case where the primary nodes also have multiple antennas will be left for further research.




(a) Time slot $t_1$




(b) Time slot $t_2$

Figure 2.2: CSI estimation at secondary node $S_1$.

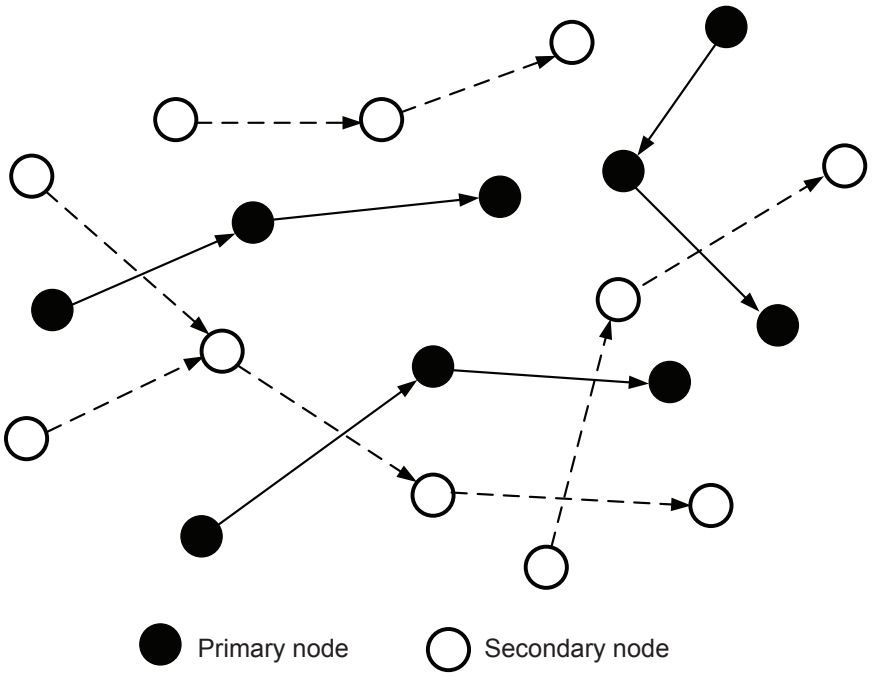



Figure 2.3: A multi-hop secondary network co-located with a multi-hop primary network.

- We further assume that the CSI obtained at the secondary nodes is perfect. This assumption allows us to develop an information theoretic understanding on the potential benefits of transparent coexistence paradigm. In practice, perfect CSI is hard to achieve and inaccurate CSI will cause interference leakage. This may be treated as additional noise and will degrade link quality. Just like any other system, there is a gap between what a theoretical limit is and what can actually be achieved in practice. Investigation of this gap (between theoretical limit and achievable performance in practice) and how to close this gap will be deferred for future research.

- We assume each data stream is associated with the same constant rate. In practice, the data rate of a data stream depends on channel condition and many other factors. But for tractability, we assume that we use a simple fixed rate coding and modulation scheme for a data stream. In other words, we assume that there is a minimum rate with our fixed rate coding and modulation for a data stream and we will just use this minimum rate for all data streams, despite that some streams with better channels could in fact achieve higher rates if

an adaptive coding and modulation scheme is used. We agree that such a simple fixed rate coding and modulation scheme is not optimal. But this assumption allows us to keep the problem tractable when performing performance study.

- In our throughput optimization problem in the transparent coexistence paradigm, we assume to have global knowledge so that we can develop a centralized solution and use it to examine the benefits of such a paradigm.

Based on these assumptions, we explore the following challenges in the secondary network:

- **Channel/time slot scheduling** In a secondary network, an intermediate relay node is both a transmitter and a receiver. Under the half-duplex, a node cannot transmit and receive on the same channel within the same time slot. Therefore, scheduling (either in time slot or channel) is needed. Here, scheduling can be performed both in time slot and channel allocation (time and frequency domains). Note that scheduling transmission/reception at a secondary node will lead to a particular interference relationship among the primary and secondary nodes in the underlying time slot and channel. This joint time/channel scheduling plays an integral role for IC in the network.

- **Inter-network IC** We discussed this challenge in Section 2.2 (see Fig. 2.1), where a secondary transmitter needs to cancel its interference to its neighboring primary receivers while a secondary receiver needs to cancel the interference from its neighboring primary transmitters.

- **Intra-network IC** In addition to inter-network IC, interference from a secondary node may also interfere with another secondary node within their own network (i.e., "intra-network" interference). Such an interference must also be canceled properly (either by a secondary transmitter or receiver) to ensure successful data communications inside the secondary network.

It is important to realize that the above three key challenges are not independent, but deeply intertwined with each other. In particular, channel/time slot scheduling at a secondary node is

directly tied to the interference relationship between the primary and secondary nodes as well as interference among the secondary nodes. Therefore, a mathematical modeling of transparent coexistence paradigm must capture all these components jointly.

## 2.4 Mathematical Modeling

In this section, we develop a mathematical model for the transparent coexistence paradigm under which a multi-hop secondary network can access the same spectrum as a primary network (see Fig. 2.3). This mathematical model will address the challenges outlined in the last section through a joint formulation.

### 2.4.1 Notation

Table 7.1 lists notation in this chapter. Suppose there is a set of sessions $\tilde{\mathcal{F}}$ within the primary network $\mathcal{P}$. For a given routing for each session, denote $\tilde{\mathcal{L}}$ as the set of links in the primary network that are traversed by these sessions (shown in solid arrow lines in Fig. 2.3). Denote $\tilde{z}^b_{(\tilde{l})}(t)$ as the number of data streams over primary link $\tilde{l} \in \tilde{\mathcal{L}}$ on channel $b$ in time slot $t$. Since a primary node only has one antenna, $\tilde{z}^b_{(\tilde{l})}(t) = 1$ if link $\tilde{l}$ is active (on channel $b$ and time slot $t$) and $0$ otherwise.

For the secondary network, we assume MIMO capability at each node. Denote $A_i$ as the number of antennas on a secondary node $i \in \mathcal{S}$. Suppose there is a set of multi-hop sessions $\mathcal{F}$ in $\mathcal{S}$. For a given routing for each session, denote $\mathcal{L}$ as the set of secondary links (shown in dashed arrow line in Fig. 2.3).

To model scheduling at a secondary node for transmission or reception, we denote $x_i^b(t)$ and $y_i^b(t)$ ($i \in \mathcal{S}, b \in \mathcal{B}$ and $1 \leq t \leq T$) as whether node $i$ is a transmitter or receiver on channel $b$ in

Table 2.1: Notation

| | |
|---|---|
| **Primary Network** | |
| $\mathcal{P}$ | The set of nodes in the primary network |
| $T$ | The number of time slots in a frame |
| $\mathcal{B}$ | The sets of channels owned by the primary network |
| $B$ | The number of channels in set $\mathcal{B}$, $B = \lvert\mathcal{B}\rvert$ |
| $\tilde{\mathcal{F}}$ | The set of sessions in the primary network |
| $\tilde{\mathcal{I}}_i$ | The set of primary nodes within the interference range of secondary node $i$ |
| $\tilde{\mathcal{L}}_i^{\mathrm{In}}$ | The set of incoming links (from other primary nodes) at node $i \in \mathcal{P}$ |
| $\tilde{\mathcal{L}}_i^{\mathrm{Out}}$ | The set of outgoing links (to other primary nodes) at node $i \in \mathcal{P}$ |
| $\tilde{\mathcal{L}}$ | The set of links in the primary network |
| $\tilde{z}_{(\tilde{l})}^b(t)$ | The number of data streams over primary link $\tilde{l}$ on channel $b$ in time slot $t$ |
| **Secondary Network** | |
| $\mathcal{S}$ | The set of nodes in the secondary network |
| $A_i$ | The number of antennas at secondary node $i \in \mathcal{S}$ |
| $\mathcal{F}$ | The set of sessions in the secondary network |
| $\mathcal{I}_i$ | The set of node in $\mathcal{S}$ that are within the interference range of secondary node $i$ |
| $\mathcal{L}_i^{\mathrm{In}}$ | The set of incoming links (from other secondary nodes) at node $i \in \mathcal{S}$ |
| $\mathcal{L}_i^{\mathrm{Out}}$ | The set of outgoing links (to other secondary nodes) at node $i \in \mathcal{S}$ |
| $\mathcal{L}$ | The set of secondary links |
| $r(f)$ | The data rate of the session $f \in \mathcal{F}$ |
| $r_{\min}$ | The minimum data rate among all secondary sessions |
| $\mathrm{Rx}(l)$ | The receiver of link $l \in \mathcal{L}$ |
| $\mathrm{Tx}(l)$ | The transmitter of link $l \in \mathcal{L}$ |
| $x_i^b(t)$ | $= 1$ if node $i \in \mathcal{S}$ is a transmitter on channel $b$ in time slot $t$, and is $0$ otherwise |
| $y_i^b(t)$ | $= 1$ if node $i \in \mathcal{S}$ is a receiver on channel $b$ in time slot $t$, and is $0$ otherwise |
| $z_{(l)}^b(t)$ | The number of data streams over link $l \in \mathcal{L}$ on channel $b$ in time slot $t$ |
| $\lambda_{j,i}^b(t)$ | The number of DoFs used by transmit node $i \in \mathcal{S}$ to cancel its interference to receive node $j \in \mathcal{S}$ on channel $b$ in time slot $t$ |
| $\mu_{j,i}^b(t)$ | The number of DoFs used by receive node $i \in \mathcal{S}$ to cancel the interference from transmit node $j \in \mathcal{S}$ on channel $b$ in time slot $t$ |
| $\theta_{j,i}^b(t)$ | Binary indicator showing the relationship between nodes $i$ and $j$ in ordered list on channel $b$ in time slot $t$ |
| $\pi^b(t)$ | An ordering for IC among the secondary nodes on channel $b$ in the time slot $t$ |
| $\pi_i^b(t)$ | The position of node $i \in \mathcal{S}$ in $\pi^b(t)$ |

time slot $t$, respectively. We have

$$x_i^b(t) = \begin{cases} 1 & \text{if node } i \text{ is a transmitter on channel } b \\ & \text{in time slot } t; \\ 0 & \text{otherwise.} \end{cases}$$

$$y_i^b(t) = \begin{cases} 1 & \text{if node } i \text{ is a receiver on channel } b \\ & \text{in time slot } t; \\ 0 & \text{otherwise.} \end{cases}$$

Under half-duplex (a node cannot transmit and receive on the same channel in the same time slot), we have the following constraint on $x_i^b(t)$ and $y_i^b(t)$:

$$x_i^b(t) + y_i^b(t) \leq 1 \qquad (i \in \mathcal{S}, b \in \mathcal{B}, 1 \leq t \leq T). \tag{2.4.1}$$

### 2.4.2 Node Ordering for IC in Secondary Network

Recall that the secondary network is solely responsible for "inter-network" IC (in addition to "intra-network" IC). To avoid unnecessary duplication in allocating DoFs for IC, it was shown in [59] that node-ordering based IC is very effective. Under this scheme, all secondary nodes are put into an ordered list. DoF allocation at each secondary node for IC is based on the position of the node in the list. It was shown in [59] that such disciplined approach can ensure: (i) there is no duplication in IC (and thus no waste of DoF resources), and (ii) the final DoF allocation is feasible. We will describe the specific rules for DoF allocation at a secondary node for IC (depending on whether it is a transmitter or receiver) in the following two sections. But first, we give a mathematical model for the node ordering concept.

Denote $\pi^b(t)$ as an ordered list of the secondary nodes in the network on $b \in \mathcal{B}$ and $1 \leq t \leq T$, and denote $\pi_i^b(t)$ as the position of node $i \in \mathcal{S}$ in $\pi^b(t)$. Therefore, $1 \leq \pi_i^b(t) \leq S$, where $S = |\mathcal{S}|$. For example, if $\pi_i^b(t) = 3$, then it means that node $i$ is the third node in the list $\pi^b(t)$.

To model the relative ordering between any two secondary nodes $i$ and $j$ in $\pi^b(t)$, we use a

binary variable $\theta^b_{j,i}(t)$ and define it as follows:

$$\theta^b_{j,i}(t) = \begin{cases} 1 & \text{if node } j \text{ is before node } i \text{ in } \pi^b(t); \\ 0 & \text{otherwise.} \end{cases}$$

It was shown in [59] that the following relationships hold among $\pi^b_i(t), \pi^b_j(t)$ and $\theta^b_{j,i}(t)$.

$$\pi^b_i(t) - S \cdot \theta^b_{j,i}(t) + 1 \leq \pi^b_j(t) \leq \pi^b_i(t) - S \cdot \theta^b_{j,i}(t) + S - 1 \,, \tag{2.4.2}$$

where $i, j \in \mathcal{S}, b \in \mathcal{B}$, and $1 \leq t \leq T$.

We point out that such a node ordering approach for DoF allocation is the most efficient approach among *all* existing DoF models that can guarantee feasibility. As pointed out in [59], an "optimal" node ordering can be found by inserting the above ordering relationship as a constraint into the overall formulation of the optimization problem, as we shall do in Section 2.5.

### 2.4.3 DoF Allocation at A Secondary Transmitter

At a secondary transmitter $i$, it needs to expend DoFs for (i) SM, (ii) IC to neighboring primary receivers, and (iii) IC to a subset of its neighboring secondary receivers based on their orders in the node list.

**(i) DoF for SM.** For SM, denote $z^b_{(l)}(t)$ and $\mathcal{L}^{\text{Out}}_i$ as the number of data streams on link $l \in \mathcal{L}$ and the set of outgoing links from secondary node $i$. Then the number of DoFs at secondary node $i \in \mathcal{S}$ for SM is $\sum_{l \in \mathcal{L}^{\text{Out}}_i} z^b_{(l)}(t)$ for $b \in \mathcal{B}$ and $1 \leq t \leq T$.

**(ii) DoF for IC to neighboring primary receivers.** To ensure transparent coexistence, a secondary transmitter needs to cancel its interference to neighboring primary receivers. Recall that if a primary receiver $p \in \mathcal{P}$ is within the interference range of node $i$, the number of DoFs at node $i$ that is used for canceling the interference to node $p$ is equal to the number of data stream that are received at node $p$. Denote $\tilde{\mathcal{L}}^{\text{In}}_p$ as the set of incoming primary links to node $p$. Denote $\tilde{\mathcal{I}}_i$ as the set of primary nodes that are located within the interference range of secondary transmitter $i$. For node

$p \in \tilde{\mathcal{I}}_i$, the number of DoFs used at node $i$ for canceling interference to node $p$ is $\sum_{\tilde{l} \in \tilde{\mathcal{L}}_p^{\text{In}}} \tilde{z}_{(\tilde{l})}^b(t)$ for $b \in \mathcal{B}$ and $1 \leq t \leq T$. Now for all primary receive nodes in $\tilde{\mathcal{I}}_i$, the number of DoFs used at node $i$ to cancel interference to these nodes is $\left(\sum_{p \in \tilde{\mathcal{I}}_i} \sum_{\tilde{l} \in \tilde{\mathcal{L}}_p^{\text{In}}} \tilde{z}_{(\tilde{l})}^b(t)\right)$ for $b \in \mathcal{B}$ and $1 \leq t \leq T$.

**(iii) DoF for IC to secondary receivers.** For IC within the secondary network, this secondary transmitter $i$ only needs to cancel its interference to a subset (instead of all) of its neighboring secondary receivers based on the node ordering list [59]. Specifically, this secondary transmitter $i$ only needs to expend DoFs to null its interference to neighboring secondary receivers that are *before* itself in the ordered secondary node list $\pi^b(t)$. Node $i$ does not need to expend any DoF to null its interference to those secondary receivers that are *after* itself in the ordered node list $\pi^b(t)$. This is because the interference from node $i$ to those secondary receivers (that are after this node in $\pi^b(t)$) will be nulled by those secondary receivers later (when we perform DoF allocation at those nodes). This is the key to avoid duplication in IC.

Recall that if a secondary receiver $j \in \mathcal{S}$ is within the interference range of secondary transmit node $i$, the number of DoFs required at transmit node $i$ to cancel its interference to node $j$ is equal to the number of data stream that are being received at node $j$. Denote $\mathcal{L}_j^{\text{In}}$ as the set of incoming links to node $j$. Denote $\mathcal{I}_i$ as the set of secondary nodes that are located within the interference range of node $i$. For secondary receive node $j \in \mathcal{I}_i$, the number of DoFs used at secondary transmit node $i$ for canceling its interference to node $j$ is $\left(\theta_{j,i}^b(t) \cdot \sum_{k \in \mathcal{L}_j^{\text{In}}}^{\text{Tx}(k) \neq i} z_{(k)}^b(t)\right)$. Note that we are using the indicator variable $\theta_{j,i}^b(t)$ to consider only those secondary receive nodes that are before node $i$ in the ordered node list $\pi^b(t)$. Now for all secondary receive nodes in $\mathcal{I}_i$, the number of DoFs used at node $i$ to cancel interference to these nodes is $\sum_{j \in \mathcal{I}_i} \left(\theta_{j,i}^b(t) \cdot \sum_{k \in \mathcal{L}_j^{\text{In}}}^{\text{Tx}(k) \neq i} z_{(k)}^b(t)\right)$ for $b \in \mathcal{B}$ and $1 \leq t \leq T$.

**Total DoF consumption.** Putting all these DoF consumptions together at a secondary transmitter $i$, we have the following constraints:

- If this secondary transmit node $i$ is active, i.e., $x_i^b(t) = 1$, we have
$$x_i^b(t) \leq \sum_{l \in \mathcal{L}_i^{\text{Out}}} z_{(l)}^b(t) + \left(\sum_{p \in \tilde{\mathcal{I}}_i} \sum_{\tilde{l} \in \tilde{\mathcal{L}}_p^{\text{In}}} \tilde{z}_{(\tilde{l})}^b(t)\right) + \sum_{j \in \mathcal{I}_i} \left(\theta_{j,i}^b(t) \cdot \sum_{k \in \mathcal{L}_j^{\text{In}}}^{\text{Tx}(k) \neq i} z_{(k)}^b(t)\right) \leq A_i \text{ ,} \quad (2.4.3)$$

which means that the DoF consumption at node $i$ cannot be more that the total number of its antennas.

- If node $i$ is not active, i.e., $x_i^b(t) = 0$, we have

$$\sum_{l \in \mathcal{L}_i^{\mathrm{Out}}} z_{(l)}^b(t) = 0 \ . \tag{2.4.4}$$

We can rewrite (2.4.3) and (2.4.4) into the following two mathematical constraints:

$$x_i^b(t) \leq \sum_{l \in \mathcal{L}_i^{\mathrm{Out}}} z_{(l)}^b(t) + \left( \sum_{p \in \tilde{\mathcal{I}}_i} \sum_{\tilde{l} \in \tilde{\mathcal{L}}_p^{\mathrm{In}}} \tilde{z}_{(\tilde{l})}^b(t) \right) +$$
$$\sum_{j \in \mathcal{I}_i} \left( \theta_{j,i}^b(t) \cdot \sum_{k \in \mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k) \neq i} z_{(k)}^b(t) \right) \leq A_i x_i^b(t) + \left(1 - x_i^b(t)\right) M, \tag{2.4.5}$$

$$\sum_{l \in \mathcal{L}_i^{\mathrm{Out}}} z_{(l)}^b(t) \leq x_i^b(t) \cdot A_i \ , \tag{2.4.6}$$

where $M$ is a large constant, which is an upper bound of $[\sum_{p \in \tilde{\mathcal{I}}_i} \sum_{\tilde{l} \in \tilde{\mathcal{L}}_p^{\mathrm{In}}} \tilde{z}_{(\tilde{l})}^b(t) + \sum_{j \in \mathcal{I}_i} \theta_{j,i}^b(t) \cdot \sum_{k \in \mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k) \neq i} z_{(k)}^b(t)]$ when $x_i^b(t) = 0$. For example, we can set $M = \sum_{j \in \mathcal{I}_i} A_j + \sum_{p \in \tilde{\mathcal{I}}_i} \sum_{\tilde{l} \in \tilde{\mathcal{L}}_p^{\mathrm{In}}} \tilde{z}_{(\tilde{l})}^b(t)$.

To see that (2.4.5) and (2.4.6) can replace (2.4.3) and (2.4.4), note that (i) when $x_i^b(t) = 1$, (2.4.5) becomes (2.4.3) and (2.4.6) holds trivially; (ii) when $x_i^b(t) = 0$, (2.4.4) and (2.4.6) are equivalent, and (2.4.5) holds trivially.

**Reformulation.** Since (2.4.5) has a nonlinear term $\left( \theta_{j,i}^b(t) \cdot \sum_{k \in \mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k) \neq i} z_{(k)}^b(t) \right)$, we can use *Reformulation-Linearization Technique* (RLT) [27, Chapter 6] to reformulate this nonlinear term by introducing new variables and adding new linear constraints. We define a new variable $\lambda_{j,i}^b(t)$ as follows:

$$\lambda_{j,i}^b(t) = \theta_{j,i}^b(t) \cdot \sum_{k \in \mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k) \neq i} z_{(k)}^b(t) \ ,$$

where $i \in \mathcal{S}, j \in \mathcal{I}_i, b \in \mathcal{B}$, and $1 \leq t \leq T$. For binary variable $\theta_{j,i}^b(t)$, we have the following

associated constraints:

$$\begin{aligned} \theta_{j,i}^{b}(t) &\geq 0\,, \\ (1-\theta_{j,i}^{b}(t)) &\geq 0\,. \end{aligned}$$

For $\sum_{k\in\mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k)\neq i} z_{(k)}^{b}(t)$, we have:

$$\begin{aligned} \sum_{k\in\mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k)\neq i} z_{(k)}^{b}(t) &\geq 0\,, \\ A_j - \sum_{k\in\mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k)\neq i} z_{(k)}^{b}(t) &\geq 0\,. \end{aligned}$$

We can cross-multiply the two constraints involving $\theta_{j,i}^{b}(t)$ with the two constraints involving $\sum_{k\in\mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k)\neq i} z_{(k)}^{b}(t)$, and replacing the product term $\left(\theta_{j,i}^{b}(t)\cdot\sum_{k\in\mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k)\neq i} z_{(k)}^{b}(t)\right)$ with $\lambda_{j,i}^{b}(t)$. Then (2.4.5) can be replaced by the following linear constraints:

$$x_i^b(t) \leq \sum_{l\in\mathcal{L}_i^{\mathrm{Out}}} z_{(l)}^{b}(t) + \left(\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{In}}} \tilde{z}_{(\tilde{l})}^{b}(t)\right) + \sum_{j\in\mathcal{I}_i}\lambda_{j,i}^{b}(t) \leq A_i x_i^b(t) + \left(1-x_i^b(t)\right)M, \quad (2.4.7)$$

$$\lambda_{j,i}^{b}(t) \geq 0, \quad (2.4.8)$$

$$\lambda_{j,i}^{b}(t) \leq \sum_{k\in\mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k)\neq i} z_{(k)}^{b}(t), \quad (2.4.9)$$

$$\lambda_{j,i}^{b}(t) \leq A_j\cdot\theta_{j,i}^{b}(t), \quad (2.4.10)$$

$$\lambda_{j,i}^{b}(t) \geq A_j\cdot\theta_{j,i}^{b}(t) - A_j + \sum_{k\in\mathcal{L}_j^{\mathrm{In}}}^{\mathrm{Tx}(k)\neq i} z_{(k)}^{b}(t), \quad (2.4.11)$$

where $i\in\mathcal{S}, j\in\mathcal{I}_i, b\in\mathcal{B}$, and $1\leq t\leq T$.

### 2.4.4 DoF Allocation at A Secondary Receiver

At a secondary receiver $i$, it needs to expend DoFs for (i) SM, (ii) canceling interference from neighboring primary transmitters, and (iii) canceling interference from a subset of its neighboring secondary transmitters based on their orders in the node list.

**(i) DoF for SM.** For SM, the number of DoFs consumed at a secondary receiver $i \in \mathcal{S}$ is $\sum_{k\in\mathcal{L}_i^{\mathrm{In}}} z_{(k)}^b(t)$ for $b \in \mathcal{B}$ and $1 \le t \le T$.

**(ii) DoF for IC from neighboring primary transmitters.** A secondary receiver needs to cancel the interference from neighboring primary transmitters. If a primary transmitter $p \in \mathcal{P}$ is within the interference range of secondary receive node $i \in \mathcal{S}$, the number of DoFs at node $i$ required for canceling this interference from node $p$ is equal to the number of data streams that are being transmitted by node $p$. Denote $\tilde{\mathcal{L}}_p^{\mathrm{Out}}$ as the set of outgoing links from primary node $p$. For $p \in \tilde{\mathcal{I}}_i$, the number of DoFs used at node $i$ for canceling interference from node $p$ is $\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{Out}}} \tilde{z}_{(\tilde{l})}^b(t)$. Now for all primary transmit nodes in $\tilde{\mathcal{I}}_i$, the number of DoFs used at node $i$ to cancel interference from these nodes is $\left(\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{Out}}} \tilde{z}_{(\tilde{l})}^b(t)\right)$ for $b \in \mathcal{B}$ and $1 \le t \le T$.

**(iii) DoF for IC from secondary transmitters.** For IC within the secondary network, this secondary receiver $i$ only needs to null the interference from a subset (instead of all) of its neighboring secondary transmitters based on node ordering list. Specifically, this secondary receiver $i$ only needs to expend DoFs to null the interference from neighboring secondary transmitters that are *before* itself in the ordered secondary node list $\pi^b(t)$. Node $i$ does not need to expend any DoF to null the interference from those secondary transmitters that are *after* itself in the ordered node list $\pi^b(t)$. This is because the interference to node $i$ from those secondary transmitters will be nulled by those secondary transmitters later (when we perform DoF allocation at those nodes).

Recall that if node $i$ is within the interference range of a secondary transmit node $j \in \mathcal{S}$, the number of DoFs at node $i$ that is used for canceling the interference from node $j$ is equal to the number of data stream that are being transmitted at node $j$. For a secondary transmit node $j \in \mathcal{I}_i$, the number of DoFs used at secondary receive node $i$ for canceling interference from node $j$ is $\left(\theta_{j,i}^b(t) \cdot \sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{Rx}(l)\neq i} z_{(l)}^b(t)\right)$. Now for all other secondary transmit nodes in $\mathcal{I}_i$, the number of DoFs used at node $i$ to cancel interference from those nodes is $\sum_{j\in\mathcal{I}_i}\left(\theta_{j,i}^b(t) \cdot \sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{Rx}(l)\neq i} z_{(l)}^b(t)\right)$ for $b \in \mathcal{B}$ and $1 \le t \le T$.

**Total DoF consumption.** We can put all DoF consumption at a secondary receiver as follows:

$$y_i^b(t) \leq \sum_{k\in\mathcal{L}_i^{\mathrm{In}}} z_{(k)}^b(t) + \left(\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{Out}}} \tilde{z}_{(\tilde{l})}^b(t)\right) + \sum_{j\in\,\mathcal{I}_i}\left(\theta_{j,i}^b(t)\cdot \sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{R}x(l)\neq i} z_{(l)}^b(t)\right) \leq A_i y_i^b(t) + \Big(1-y_i^b(t)\Big)\,N \quad (2.4.12)$$

$$\sum_{k\in\mathcal{L}_i^{\mathrm{In}}} z_{(k)}^b(t) \leq y_i^b(t)\cdot A_i\ , \quad (2.4.13)$$

where $N$ is a large constant, which is an upper bound of $[\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{Out}}}\tilde{z}_{(\tilde{l})}^b(t) + \sum_{j\in\,\mathcal{I}_i}(\theta_{j,i}^b(t)\cdot \sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{R}x(l)\neq i} z_{(l)}^b(t)]$ when $y_i^b(t)=0$. For example, we can set $N=\sum_{j\in\mathcal{I}_i}A_j+\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{Out}}}\tilde{z}_{(\tilde{l})}^b(t)$.

**Reformulation.** Following the same token as in the last section, we use RLT to linearize the nonlinear term $\left(\theta_{j,i}^b(t)\cdot\sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{R}x(l)\neq i} z_{(l)}^b(t)\right)$ in (2.4.12). Denote $\mu_{j,i}^b(t)$ as $\left(\theta_{j,i}^b(t)\cdot\sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{R}x(l)\neq i} z_{(l)}^b(t)\right)$. Then (2.4.12) can be replaced by the following linear constraints:

$$y_i^b(t) \leq \sum_{k\in\mathcal{L}_i^{\mathrm{In}}} z_{(k)}^b(t) + \left(\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\mathcal{L}_p^{\mathrm{Out}}} \tilde{z}_{(\tilde{l})}^b(t)\right) + \sum_{j\in\mathcal{I}_i}\mu_{j,i}^b(t) \leq A_i y_i^b(t) + \Big(1-y_i^b(t)\Big)\,N, \quad (2.4.14)$$

$$\mu_{j,i}^b(t) \geq 0, \quad (2.4.15)$$

$$\mu_{j,i}^b(t) \leq \sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{R}x(l)\neq i} z_{(l)}^b(t), \quad (2.4.16)$$

$$\mu_{j,i}^b(t) \leq A_j\cdot\theta_{j,i}^b(t), \quad (2.4.17)$$

$$\mu_{j,i}^b(t) \geq A_j\cdot\theta_{j,i}^b(t) - A_j + \sum_{l\in\mathcal{L}_j^{\mathrm{Out}}}^{\mathrm{R}x(l)\neq i} z_{(l)}^b(t), \quad (2.4.18)$$

where $i\in\mathcal{S}, j\in\mathcal{I}_i, b\in\mathcal{B}$, and $1\leq t\leq T$.

## 2.5 Case Study: A Throughput Maximization Problem

### 2.5.1 Problem Formulation

Using the above mathematical model for the transparent coexistence paradigm for a multi-hop secondary network, various problems can be studied. In this section, we study a throughput opti-

mization problem in the secondary network. Denote $r(f)$ as the rate of session $f \in \mathcal{F}$. Then at any link $l \in \mathcal{L}$ in the network, the aggregate throughput rate among the flows that traverse this link cannot exceed the link's scheduling capacity (over a time frame). That is,

$$\sum_{f \in \mathcal{F}}^{f \text{ traversing } l} r(f) \leq c \cdot \frac{1}{T} \sum_{b \in \mathcal{B}} \sum_{t=1}^{T} z_{(l)}^{b}(t) \qquad (l \in \mathcal{L}), \tag{2.5.1}$$

where $c$ is the data rate carried by a data stream.

For the throughput maximization problem, suppose we are interested in maximizing the minimum throughput rate among all secondary sessions. Then the problem can be formulated as follows:

OPT

max $r_{\min}$

s.t $r_{\min} \leq r(f) \qquad (f \in \mathcal{F})$;

Half-duplex constraints: (2.4.1);

Node ordering constraints: (2.4.2);

Transmitter DoF constraints: (2.4.6)–(2.4.11);

Receiver DoF constraints: (2.4.13)–(2.4.18);

Link capacity constraints: (2.5.1).

In this formulation, $r_{\min}, r(f), x_i^b(t), y_i^b(t), z_{(l)}^b(t), \pi_i^b(t)$, $\lambda_{j,i}^b(t)$, $\mu_{j,i}^b(t)$ and $\theta_{j,i}^b(t)$ are optimization variables, and $A_i, M, N, \tilde{z}_{(l)}^b(t)$ and $c$ are given constants. This optimization problem is in the form of a mixed-integer linear program (MILP), which is NP-hard in general. Although commercial solvers such as CPLEX can be used, they are not scalable to address problems with moderate to large-sized networks. In this section, we develop an efficient heuristic algorithm.

### 2.5.2 Overview of Solution Algorithm

The algorithm that we propose is based on the so-called *sequential fixing* (SF) technique in [27, Chapter 5]. SF offers a general framework to handle integer variables in a MILP problem, and has

a polynomial time complexity. The basic idea of SF is as follows. For a MILP like ours, if we were able to set the optimal values for all integer variables, then the original problem would be reduced to an LP, which can be solved in polynomial time. So the key challenge in MILP is how to determine the values for all the integer variables. Under SF, this can be done by studying the linear relaxation of the original problem, obtained by relaxing all the integer variables to continuous variables. Although the solution to this linear relaxation may not have an integer value for each integer variable, we can *fix* the values of one or more integer variables based on their *closeness* to certain integer values. Instead of determining all the integer variables in one iteration, we can fix only one or a few integer variables in each iteration. For the remaining (unfixed) integer variables, we can solve a new linear relaxation and then fix one or more remaining integer variables. This SF procedure terminates after all integer variables are fixed. At this point, the MILP becomes an LP. Any remaining continuous variable in the LP can be solved efficiently.

Although the idea of SF is straightforward, it requires a careful design to ensure its performance. A naive application of SF, as we have experienced, may lead to either infeasible solution or poor performance. This is because that fixing relaxed variables solely based on their closeness to integers do not take into consideration of the physical significance of different variables in the particular problem and their intricate relationships. In our design, we propose to classify integer variables into three groups: $(\pi, \theta)$, $(x, y)$, and $z$. The first group $(\pi, \theta)$, determines the ordering among the secondary nodes in DoF allocation and is considered the structural foundation of all integer variables. Therefore, we will determine $(\pi, \theta)$ first in our SF algorithm. For the remaining $(x, y)$ and $z$ variables, $(x, y)$ can be determined if we know the link status for the corresponding $z$. Therefore, we will determine the link status (i.e., whether $z = 0$ or $z \geq 1$) first and then we can fix the corresponding $(x, y)$. Note that in this step, we only determine whether $z = 0$ (link inactive) or $z \geq 1$ (link active). In the last step, we will fix those $z$'s with $z \geq 1$ to exact integer values iteratively. Some important details of each step are given in the following section.

### 2.5.3 Algorithm Details

**Phase I: Fixing $\pi$ and $\theta$ variables.** In this phase, for $b \in \mathcal{B}$ and $1 \leq t \leq T$, we will fix one $\pi_i^b(t)$ variable, and further fix related $\theta_{i,j}^b(t)$ (or $\theta_{j,i}^b(t)$) variables during an iteration. Since there are a total of $S$ of $\pi_i^b(t)$'s ($i \in \mathcal{S}$) for $b \in \mathcal{B}$ and $1 \leq t \leq T$, there are $S$ iterations in Phase I.

Specifically, in the first iteration, for $b \in \mathcal{B}$ and $1 \leq t \leq T$, we identify node $i$ with the smallest value of $\pi_i^b(t)$ among all $\pi_j^b(t)$'s ($j \in \mathcal{S}$). We set $\pi_i^b(t) = 1$. Since this is the first node on channel $b$ in time slot $t$, we set $\theta_{i,j}^b(t) = 1$ and $\theta_{j,i}^b(t) = 0$ for $j \neq i$. In the second iteration, another node $k$ with the smallest value $\pi_k^b(t)$ among all un-fixed $\pi_j^b(t)$'s will be chosen and we set $\pi_k^b(t) = 2$. Likewise, we set $\theta_{k,j}^b(t) = 1$ and $\theta_{j,k}^b(t) = 0$ for $j \neq i$, $j \neq k$. This process continues till the end of $S$-th iteration, when all $\pi_i^b(t)$ and $\theta_{i,j}^b(t)$ ($i, j \in \mathcal{S}$) are fixed for $b \in \mathcal{B}, 1 \leq t \leq T$.

**Phase II: Fixing $x$ and $y$ variables.** In this phase, we will determine each link $l$'s status (i.e., active or inactive) and fix $x_i^b(t)$ and $y_i^b(t)$ variables. In the case of an inactive link $l$, we set $z_{(l)}^b(t) = 0$; in the case of an active link $l$, we will leave the determination of $z_{(l)}^b(t)$ to Phase III.

Specifically, in each iteration, we choose the largest $z_{(l)}^b(t)$ on channel $b$ in time slot $t$ and determine the status of the corresponding link $l$ (i.e., active or inactive). This link $l$ is determined to be active for $b \in \mathcal{B}$ and $1 \leq t \leq T$ if it satisfies the following conditions:

- (2.4.1) is satisfied, which means that the transmitter and receiver of this link each meets half-duplex constraint.
- Link $l$'s transmitter should satisfy (2.4.5) and its receiver should satisfy (2.4.12), i.e., not exceeding DoF resources at both transmitter and receiver. In the case that the status of another associated link $k$ is yet to be determined, we assume its $z_{(k)}^b(t) = 0$. Similarly, in the case that the status of another associated link $k$ is active, we assume $z_{(k)}^b(t) = 1$. Note that in either case, we do not set the values for these $z_{(k)}^b(t)$'s permanently, but rather, only a lower bound value so that we can test whether (2.4.5) and (2.4.12) can hold.

If link $l$ does not meet the above two conditions, it is considered inactive. Depending on

whether link $l$ is active or inactive, we can fix $(x_i^b(t), y_i^b(t))$ and possibly some other $z_{(k)}^b(t)$ variables based on the following three rules:

(a) If link $l$ is active for $b \in \mathcal{B}$ and $1 \leq t \leq T$, we can fix $x_{\mathrm{Tx}(l)}^b(t) = 1$ and $y_{\mathrm{Rx}(l)}^b(t) = 1$. As a result of this fixing, we can also fix $y_{\mathrm{Tx}(l)}^b(t) = 0$ and $x_{\mathrm{Rx}(l)}^b(t) = 0$ by (2.4.1). Otherwise (i.e., link $l$ is inactive for $b \in \mathcal{B}$ and $1 \leq t \leq T$), we can fix $z_{(l)}^b(t) = 0$. Further, if all links in $\mathcal{L}_{\mathrm{Tx}(l)}^{\mathrm{Out}}$ are inactive for $b \in \mathcal{B}$ and $1 \leq t \leq T$, we set $x_{\mathrm{Tx}(l)}^b(t) = 0$. Similarly, if all links in $\mathcal{L}_{\mathrm{Rx}(l)}^{\mathrm{In}}$ are inactive for $b \in \mathcal{B}$ and $1 \leq t \leq T$, we set $y_{\mathrm{Rx}(l)}^b(t) = 0$.

(b) If $x_i^b(t) = 0$, i.e., node $i$ does not transmit data for $b \in \mathcal{B}$ and $1 \leq t \leq T$, then we set all links $k \in \mathcal{L}_i^{\mathrm{Out}}$ to be inactive. Further, we set $z_{(k)}^b(t) = 0$ on these links.

(c) If $y_i^b(t) = 0$, i.e., node $i$ does not receive data for $b \in \mathcal{B}$ and $1 \leq t \leq T$, then we set all link $k \in \mathcal{L}_i^{\mathrm{In}}$ to be inactive. Further, we set $z_{(k)}^b(t) = 0$ on these links.

We use an example to illustrate the case when a link is determined to be active. Referring to Fig. 2.4, suppose the status on links 6, 8, and 10 are determined to be inactive on $b$ and $t$ in the last iteration. In this iteration, suppose link 1's status is determined to be active. Then, we can set $x_{\mathrm{Tx}(1)}^b(t) = 1$ and $y_{\mathrm{Rx}(1)}^b(t) = 1$. Since $x_{\mathrm{Tx}(1)}^b(t) = 1$, we can set $y_{\mathrm{Tx}(1)}^b(t) = 0$ and $z_{(2)}^b(t) = 0$, $z_{(3)}^b(t) = 0$. The link status of 2 and 3 can be set to be inactive. Since all outgoing links from node $\mathrm{Tx}(3)$ are inactive, we can set $x_{\mathrm{Tx}(3)}^b(t) = 0$. Similarly, since $y_{\mathrm{Rx}(1)}^b(t) = 1$, we can set $x_{\mathrm{Rx}(1)}^b(t) = 0$ and $z_{(4)}^b(t) = 0$, $z_{(5)}^b(t) = 0$. The link status of 4 and 5 can be set to be inactive. Since all incoming links to $\mathrm{Rx}(4)$ are inactive, we can set $x_{\mathrm{Rx}(4)}^b(t) = 0$.

**Phase III: Fixing $z$ variables.** In Phase II, we have fixed $z_{(l)}^b(t)$'s to 0 for those inactive links. For those links that are active, we have not yet determined the exact integer values for $z_{(l)}^b(t)$'s. In Phase III, we will fix these integer values.

On all active links $l$, if there exists some $z_{(l)}^b(t)$'s that are not yet integer, we use SF to fix these $z_{(l)}^b(t)$'s iteratively until they are all integers. In particular, during each iteration, we identify link $l$ with the $\min_l \{z_{(l)}^b(t) - \lfloor z_{(l)}^b(t) \rfloor\}$ for each for $b \in \mathcal{B}$ and $1 \leq t \leq T$ and set $z_{(l)}^b(t) = \lfloor z_{(l)}^b(t) \rfloor$.

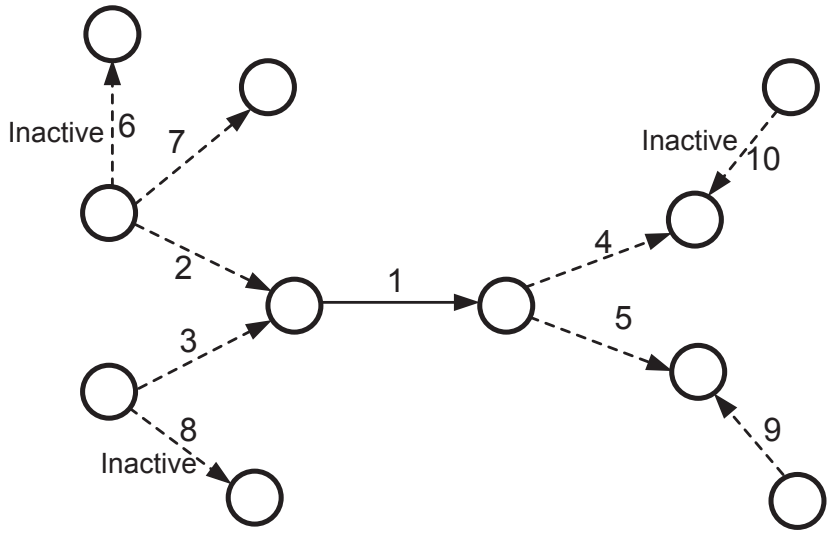



Figure 2.4: An example illustrating how to fix $x$, $y$, and some $z$ variables in Phase II.

## 2.6 Performance Evaluation

The goal of this section is twofold. First, we want to use numerical results to illustrate how transparent coexistence can be achieved for a multi-hop secondary network. Note that we cannot compare our heuristic solution to the global optimal solution because a global optimal solution is not available due to the exponential complexity of an MILP formulation. But this limitation does not prevent us from demonstrating the potential benefits of the transparent coexistence paradigm. Therefore, our second goal in this section is to show the tremendous benefits (in terms of spectrum access and throughput gain) of the transparent coexistent paradigm over the existing interference avoidance paradigm.

### 2.6.1 An Example

Consider a 20-node primary network and a 30-node secondary network randomly deployed in the same $100 \times 100$ area. For the ease of scalability and generality, we normalize all units for distance, bandwidth, and throughput with appropriate dimensions. The location for each node (both primary and secondary) is generated at random and is listed in Table 2.2. We assume that there are four

Table 2.2: Location of each node for the 20-node primary network and 30-node secondary network.

| Primary Network | | | | | |
|---|---|---|---|---|---|
| Node | Location | Node | Location | Node | Location |
| $P_1$ | (10, 10) | $P_8$ | (15, 50) | $P_{15}$ | (20, 80) |
| $P_2$ | (30, 30) | $P_9$ | (40, 70) | $P_{16}$ | (31, 48) |
| $P_3$ | (50, 30) | $P_{10}$ | (60, 90) | $P_{17}$ | (35, 85) |
| $P_4$ | (75, 50) | $P_{11}$ | (85, 90) | $P_{18}$ | (90, 80) |
| $P_5$ | (90, 20) | $P_{12}$ | (40, 10) | $P_{19}$ | (3, 35) |
| $P_6$ | (90, 45) | $P_{13}$ | (70,10) | $P_{20}$ | (6, 97) |
| $P_7$ | (75, 65) | $P_{14}$ | (55, 55) | | |
| Secondary Network | | | | | |
| Node | Location | Node | Location | Node | Location |
| $S_1$ | (23, 66) | $S_{11}$ | (55, 60) | $S_{21}$ | (88, 62) |
| $S_2$ | (3, 89) | $S_{12}$ | (8, 56) | $S_{22}$ | (70, 20) |
| $S_3$ | (42, 41) | $S_{13}$ | (3, 78) | $S_{23}$ | (76, 74) |
| $S_4$ | (19, 37) | $S_{14}$ | (62, 2) | $S_{24}$ | (84, 30) |
| $S_5$ | (10, 70) | $S_{15}$ | (92, 92) | $S_{25}$ | (22, 92) |
| $S_6$ | (29, 6) | $S_{16}$ | (36, 94) | $S_{26}$ | (60, 40) |
| $S_7$ | (8, 25) | $S_{17}$ | (82, 4) | $S_{27}$ | (28, 16) |
| $S_8$ | (51, 10) | $S_{18}$ | (35, 60) | $S_{28}$ | (99, 3) |
| $S_9$ | (63, 75) | $S_{19}$ | (76, 40) | $S_{29}$ | (98, 38) |
| $S_{10}$ | (65, 98) | $S_{20}$ | (48, 21) | $S_{30}$ | (47, 85) |

antennas on each secondary node, and all nodes' transmission range and interference range are 30 and 50, respectively.[3] There are two channels owned by the primary network ($B = 2$). A time frame is divided into four time slots ($T = 4$). For simplicity, we assume the data rate of one data stream in a time slot is 1 unit ($c = 1$).

We assume there are three active sessions in the primary network and four active sessions in the secondary network (see Table 2.3). For simplicity, we assume that minimum-hop routing is used for the primary and secondary sessions, although other routing methods will also work here.

[3]For an indepth study on how to set interference range, we refer readers to our previous work in [60].

Table 2.3: Source and destination nodes of each session in the primary and secondary networks.

| Primary Network | | |
|---|---|---|
| Session | Source Node | Destination Node |
| 1 | $P_1$ | $P_{14}$ |
| 2 | $P_5$ | $P_7$ |
| 3 | $P_{11}$ | $P_{15}$ |
| Secondary Network | | |
| Session | Source Node | Destination Node |
| 1 | $S_7$ | $S_{25}$ |
| 2 | $S_{21}$ | $S_{17}$ |
| 3 | $S_{14}$ | $S_3$ |
| 4 | $S_{30}$ | $S_{23}$ |

Further, the channel and time slot allocation on each hop for each primary session is known *a priori* and is shown in Fig. 2.5, where $(b, t)$ means this link is transmitting on channel $b$ in time slot $t$. The solid arrows represent the links in the primary network, while the dashed arrows represent the links in the secondary network.

For this network setting, we apply our solution algorithm to solve OPT. The obtained objective value is $1.0$. The channel and time slot scheduling on each link for each secondary session is shown in the shaded box as in Fig. 2.6, where $(b, t)$ on each secondary link represents that this link transmits on channel $b$ in time slot $t$. The details of DoFs used for SM on each channel in each time slot on each link in the secondary network are shown in Table 2.4. The link rate (i.e., total number of DoFs used for SM averaged over a 4-time-slot frame) on a link is also shown in this table.

To see how the secondary node can be active simultaneously with the primary nodes while remain transparent, consider $(b, t) = (1, 2)$ (channel 1, time slot 2) in Fig. 2.6. Here, link $P_3 \rightarrow P_{14}$ in the primary network is active; links $S_{14} \rightarrow S_{20}$, $S_{22} \rightarrow S_{17}$, $S_{21} \rightarrow S_{19}$, $S_{30} \rightarrow S_9$ and $S_4 \rightarrow S_1$ in the secondary network are also active. Based on a node's interference range, the interference relationships among the nodes associated with these active links are shown in Fig. 2.7, where the dotted arrow lines show the interference from a (primary or secondary) transmitter to an unintended

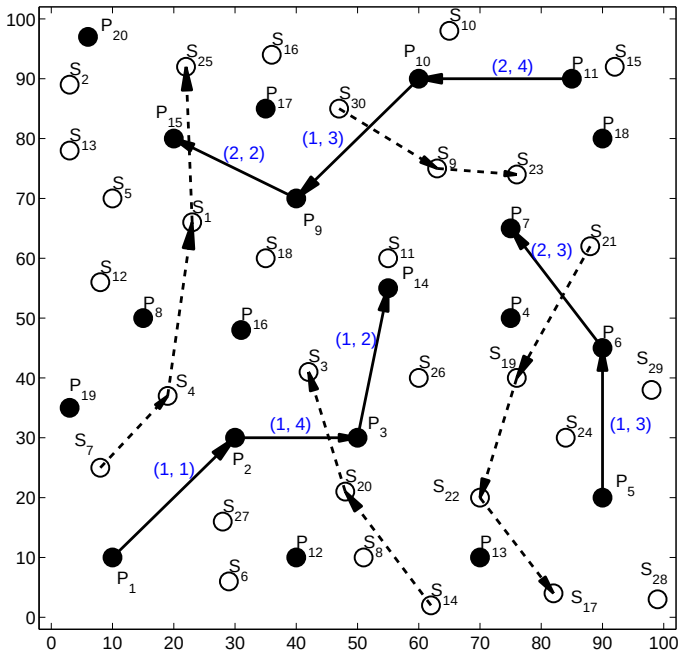



Figure 2.5: Active sessions in the primary and secondary networks.

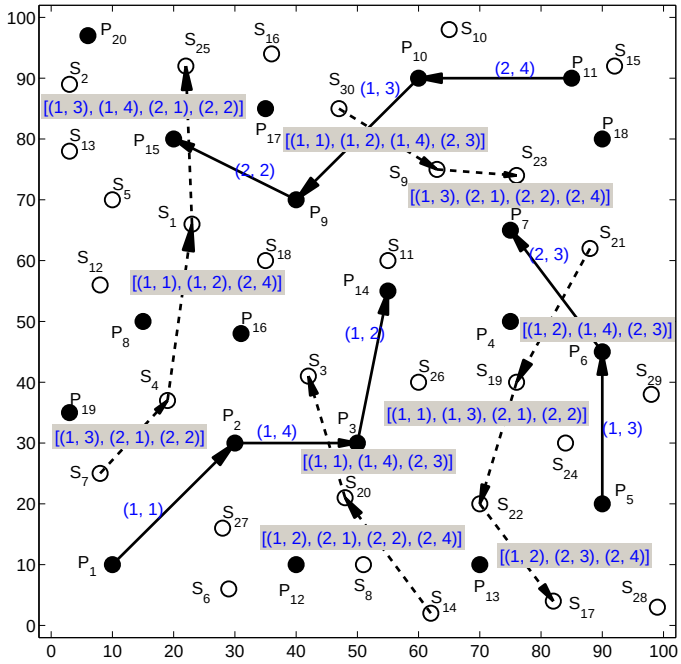



Figure 2.6: Channel and time slot scheduling on each link for the secondary sessions by our solution algorithm. Channel and time slot scheduling on each link for the primary sessions are given in Fig. 2.5.

Table 2.4: Channel and time slot scheduling on each link, DoF allocation for SM, and throughput on each link for the secondary sessions.

| Session | Link | (channel, time slot) scheduling | DoF for SM | Link rate |
|---|---|---|---|---|
| 1 | $S_7 \longrightarrow S_4$ | (1, 3) | 2 | 1.0 |
| | | (2, 1) | 1. | |
| | | (2, 2) | 1 | |
| | $S_4 \longrightarrow S_1$ | (1, 1) | 1 | 1.0 |
| | | (1, 2) | 1 | |
| | | (2, 4) | 2 | |
| | $S_1 \longrightarrow S_{25}$ | (1, 3) | 1 | 1.0 |
| | | (1, 4) | 1 | |
| | | (2, 1) | 1 | |
| | | (2, 2) | 1 | |
| 2 | $S_{21} \longrightarrow S_{19}$ | (1, 2) | 1 | 1.0 |
| | | (1, 4) | 2 | |
| | | (2, 3) | 1 | |
| | $S_{19} \longrightarrow S_{22}$ | (1, 1) | 1 | 1.0 |
| | | (1, 3) | 1 | |
| | | (2, 1) | 1 | |
| | | (2, 2) | 1 | |
| | $S_{22} \longrightarrow S_{17}$ | (1, 2) | 1 | 1.0 |
| | | (2, 3) | 1 | |
| | | (2, 4) | 2 | |
| 3 | $S_{14} \longrightarrow S_{20}$ | (1, 2) | 1 | 1.0 |
| | | (2, 1) | 1 | |
| | | (2, 2) | 1 | |
| | | (2, 4) | 1 | |
| | $S_{20} \longrightarrow S_3$ | (1, 1) | 2 | 1.0 |
| | | (1, 4) | 2 | |
| | | (2, 3) | 1 | |
| 4 | $S_{30} \longrightarrow S_9$ | (1, 1) | 1 | 1.0 |
| | | (1, 2) | 1 | |
| | | (1, 4) | 1 | |
| | | (2, 3) | 1 | |
| | $S_9 \longrightarrow S_{23}$ | (1, 3) | 1 | 1.0 |
| | | (2, 1) | 1 | |
| | | (2, 2) | 1 | |
| | | (2, 4) | 1 | |

Table 2.5: DOF allocation for SM and IC on $(b, t) = (1, 2)$ at each node in the secondary network.

| Node $i$ | TX/RX | $\pi_i^1(2)$ | DoF for SM | DoF for IC to/from primary nodes | DoF for IC within secondary network |
|---|---|---|---|---|---|
| $S_{19}$ | RX | 1 | 1 | 1 from $P_3$ | 0 |
| $S_{14}$ | TX | 2 | 1 | 0 | 1 to $S_{19}$ |
| $S_{22}$ | TX | 4 | 1 | 1 to $P_{14}$ | 1 to $S_{19}$ |
| $S_{21}$ | TX | 5 | 1 | 1 to $P_{14}$ | 0 |
| $S_{17}$ | RX | 6 | 1 | 1 from $P_3$ | 1 from $S_{14}$ |
| $S_{20}$ | RX | 8 | 1 | 1 from $P_3$ | 1 from $S_{22}$ |
| $S_{30}$ | TX | 9 | 1 | 1 to $P_{14}$ | 0 |
| $S_9$ | RX | 11 | 1 | 1 from $P_3$ | 1 from $S_{21}$ |
| $S_4$ | TX | 12 | 1 | 1 to $P_{14}$ | 1 to $S_{20}$ |
| $S_1$ | RX | 13 | 1 | 1 from $P_3$ | 1 from $S_{30}$ |

(primary or secondary) receiver. Table 2.5 shows the DoF allocation at each secondary node for SM, IC to/from primary nodes, and IC within the secondary network for $(b, t) = (1, 2)$.

- First, we check whether there is any interference to primary receiver $P_{14}$. Note that there are four potential interference from secondary transmitters, i.e., $S_4$, $S_{21}$, $S_{22}$ and $S_{30}$. Since each of these secondary transmitter uses one DoF to cancel its interference to primary receiver $P_{14}$ (fifth column in Table 2.5), all interference on the primary receiver $P_{14}$ is effectively nulled. Therefore, the primary receiver $P_{14}$ is not interfered by the simultaneous activation of its neighboring secondary transmitters.

- Next, we check whether the interference from the primary transmitter is nulled properly at its neighboring secondary receivers ("inter-network" interference). Note that primary transmit node $P_3$ is interfering its neighboring secondary receive nodes $S_1$, $S_{20}$, $S_{17}$, $S_{19}$ and $S_9$. Since each of these secondary receive nodes uses one DoF to cancel this interference (fifth column in Table 2.5), this interference from primary transmit node $P_3$ is effectively nulled at these secondary receive nodes.

- Finally, we check whether the interference within the secondary network ("intra-network" interference) is nulled properly by the secondary nodes themselves. The IC within the secondary network follows the node ordering, which is shown in the third column of Table 2.5. The number of DoFs used for IC to/from other secondary nodes is shown in the last column of Table 2.5. As an example, consider node $S_{22}$, which is a transmit node. Referring to Table 2.5, $S_{22}$ only needs to cancel its interference to those receive nodes that are before itself in the ordered node list and within $S_{22}$'s interference range, i.e., node $S_{19}$. Table 2.5 (last column) shows that $S_{22}$ indeed uses one DoF to cancel its interference to $S_{19}$. For its interference to the secondary receive node $S_{20}$ which is also in $S_{22}$'s interference range, $S_{22}$ does not need to do anything as $S_{20}$ is after node $S_{22}$ in the ordered list. This interference to $S_{20}$ will be canceled by $S_{20}$ (as shown in Table 2.5, last column).

  It can be easily verified that for all interference among the active secondary nodes are properly canceled. Further, at each active secondary node, the DoFs used for SM, IC to/from the primary nodes, IC within the secondary network is not more than its total DoFs (i.e., 4).

The above illustration is for $(b, t) = (1, 2)$ (i.e., channel 1, time slot 2), the results for the other channel and time slots (i.e., $(1, 1)$, $(1, 4)$, $(1, 3)$, $(2, 2)$, $(2, 3)$ and $(2, 4)$) are similar and are omitted to conserve space.

### 2.6.2 Comparison to Interference Avoidance Paradigm

To see the benefits of the transparent coexistence paradigm, we compare it to the prevailing interference avoidance paradigm. Under the interference avoidance paradigm, a secondary node is not allowed to transmit (receive) on the same channel at the same time when a nearby primary receiver (transmitter) is using this channel. Therefore, the set of available channel and time slots that can be used by secondary nodes is smaller. The problem formulation for this paradigm is similar to (but simpler than) OPT. In particular, we can remove the second term ($\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{In}}}\tilde{z}^b_{(\tilde{l})}(t)$ and $\sum_{p\in\tilde{\mathcal{I}}_i}\sum_{\tilde{l}\in\tilde{\mathcal{L}}_p^{\mathrm{Out}}}\tilde{z}^b_{(\tilde{l})}$ ) in constraints (2.4.5) and (2.4.12) in OPT that are used for secondary nodes

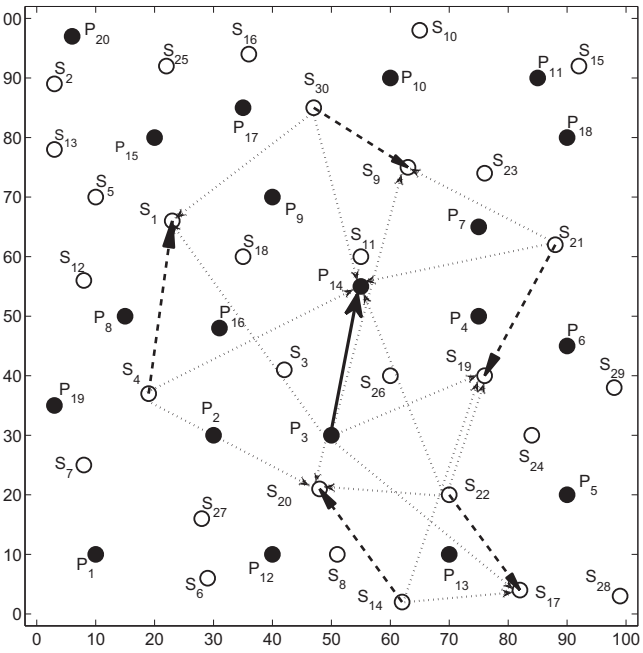



Figure 2.7: Illustration of interference relationships among the primary and secondary links on channel 1 in time slot 2 in the case study.

to cancel interference to/from the primary nodes. The problem formulation remains an MILP and a solution algorithm similar to that in Section 2.5.3 can be used to solve it.

Following the same setting as in the case study in Section 2.6.1, we solve the above optimization problem under the interference avoidance paradigm. Note that the available channels and time slot resources at each node are only a subset of 2 channels and 4 time slots, versus full 2 channels and 4 time slots for each secondary node in the transparent coexistence paradigm. The obtained objective value is $0.5$ (compared to $1.0$ in Section 2.6.1). The channel and time slot scheduling on each link of each secondary session is shown in Fig. 2.8. Comparing Figs. 2.6 and 2.8, we find that the set of channels and time slots used by each secondary link under interference avoidance paradigm is smaller. The details for the DoF allocation for SM on each channel in each time slot and link rate are shown in Table 2.6. Comparing Tables 2.6 and 2.4, the rates on most links are smaller under the interference avoidance paradigm.

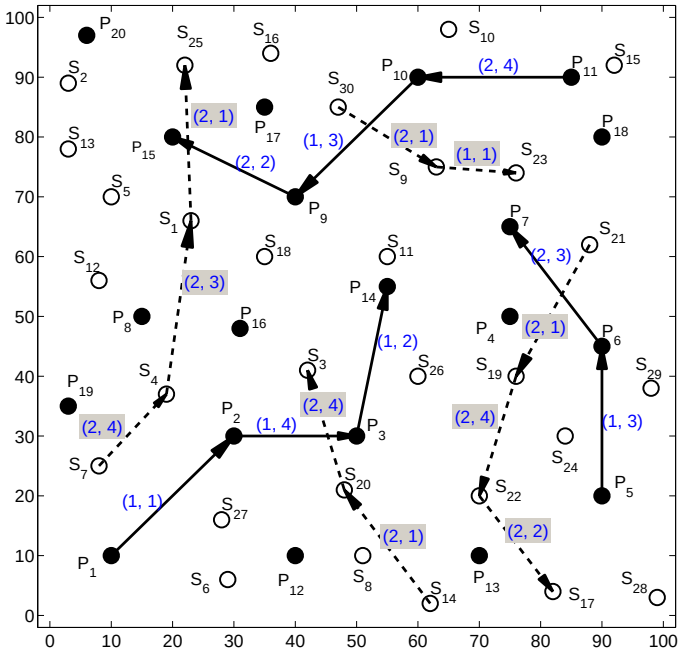
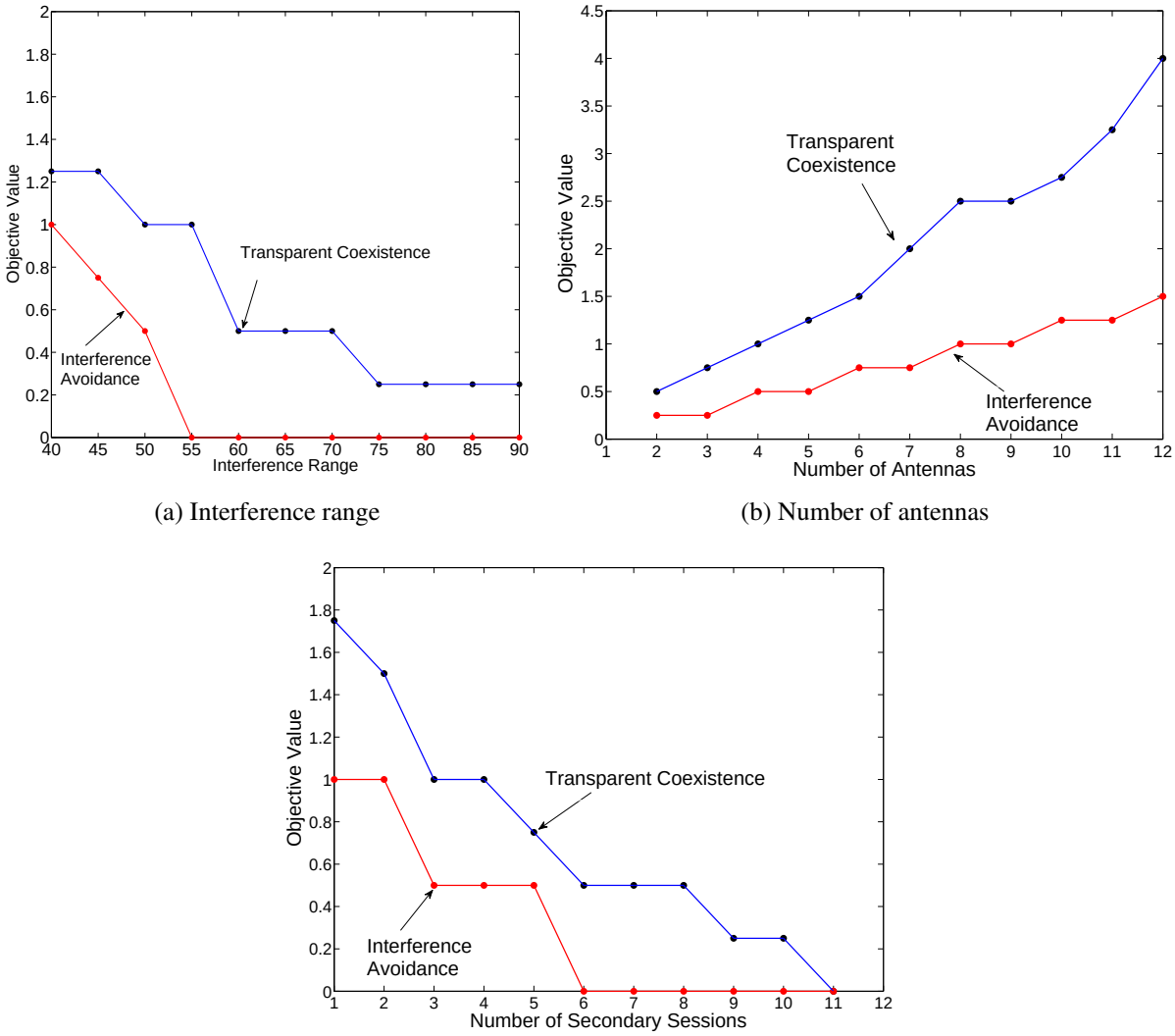



Figure 2.8: Channel and time slot scheduling on each link for the secondary sessions under the interference avoidance paradigm.

### 2.6.3 Impact of Various System Parameters

The results in Sections 2.6.1 and 2.6.2 show the solution details for a case study in the transparent coexistence paradigm and its improvement in objective value over that in the interference avoidance paradigm. To show the robustness of our results, we further perform numerical study for the same network under different system parameters, such as interference range setting, the number of antennas on each node, and the number of sessions in the secondary network.

Fig. 2.9(a) shows the objective values under the transparent coexistence paradigm and the interference avoidance paradigm when the interference range for the secondary network is varied from 40 to 90 (while keeping the transmission range at 30). As shown in the figure, the performance under the transparent coexistence paradigm is always better than that under the interference avoidance paradigm for the same interference range, although the performance under both paradigms degrades when the interference range increases.




(a) Interference range

(b) Number of antennas

(c) Number of secondary sessions

Figure 2.9: Impact of the various system parameters on the performance of transparent coexistence and interference avoidance paradigms.

Table 2.6: Channel and time slot scheduling on each link, DoF allocation for SM, and link rate on each link for the secondary sessions under the interference avoidance paradigm.

| Session | Link | (channel, time slot) scheduling | DoF for SM | Link rate |
|---|---|---|---|---|
| 1 | $S_7 \longrightarrow S_4$ | (2, 4) | 2 | 0.5 |
| | $S_4 \longrightarrow S_1$ | (2, 3) | 4 | 1.0 |
| | $S_1 \longrightarrow S_{25}$ | (2, 1) | 2 | 0.5 |
| 2 | $S_{21} \longrightarrow S_{19}$ | (2, 1) | 2 | 0.5 |
| | $S_{19} \longrightarrow S_{22}$ | (2, 4) | 2 | 0.5 |
| | $S_{22} \longrightarrow S_{17}$ | (2, 2) | 4 | 1.0 |
| 3 | $S_{14} \longrightarrow S_{20}$ | (2, 1) | 2 | 0.5 |
| | $S_{20} \longrightarrow S_3$ | (2, 4) | 2 | 0.5 |
| 4 | $S_{30} \longrightarrow S_9$ | (2, 1) | 2 | 0.5 |
| | $S_9 \longrightarrow S_{23}$ | (1, 1) | 4 | 1.0 |

Fig. 2.9(b) shows the comparison of objective values with different antenna numbers for each secondary node under the two paradigms. Interference range for the secondary nodes is set to 50. For MIMO, the minimum number of antennas on a node is 2. As shown in the figure, the objective value under the transparent coexistence paradigm is always better than that under the interference avoidance paradigm for the same number of antennas. Further, the objective value increases under both paradigms.

Fig. 2.9(c) shows the comparison of objective values with different number of secondary sessions under the two paradigms. The number of antennas on each secondary node is 4. As shown in the figure, the objective value under the transparent coexistence paradigm is always better than that under the interference avoidance paradigm for the same number of secondary sessions, although the objective value decreases under both paradigms when the number of secondary sessions increases.

Table 2.7: Achievable minimum session throughput under transparent coexistence paradigm and interference avoidance paradigm for 50 cases.

| Network Instance | Transparent Coexistence | Interference Avoidance | Improvement | Network Instance | Transparent Coexistence | Interference Avoidance | Improvement |
|---|---|---|---|---|---|---|---|
| 1 | 1.0 | 0.5 | 100% | 26 | 0.5 | 0 | $\infty$ |
| 2 | 1.0 | 0.5 | 100% | 27 | 0.75 | 0.5 | 50% |
| 3 | 1.25 | 0.75 | 66.7% | 28 | 1.0 | 0.5 | 100% |
| 4 | 1.0 | 0.5 | 100% | 29 | 0.25 | 0 | $\infty$ |
| 5 | 1.0 | 0 | $\infty$ | 30 | 1.0 | 0.75 | 33.3% |
| 6 | 1.0 | 0.75 | 33.3% | 31 | 1.5 | 0.75 | 100% |
| 7 | 1.0 | 0 | $\infty$ | 32 | 1.25 | 0 | $\infty$ |
| 8 | 1.0 | 0.5 | 100% | 33 | 1.0 | 0.5 | 100% |
| 9 | 1.5 | 1 | 50% | 34 | 1.0 | 0.5 | 100% |
| 10 | 1.0 | 0.5 | 50% | 35 | 1.25 | 0.75 | 66.7% |
| 11 | 1.0 | 0.5 | 50% | 36 | 0.75 | 0.5 | 50% |
| 12 | 1.0 | 0.75 | 33.3% | 37 | 0.5 | 0 | $\infty$ |
| 13 | 1.25 | 0.75 | 66.7% | 38 | 1.0 | 0.25 | 300% |
| 14 | 1.0 | 0 | $\infty$ | 39 | 0.25 | 0 | $\infty$ |
| 15 | 1.0 | 0.5 | 100% | 40 | 1.0 | 0.5 | 100% |
| 16 | 1.0 | 0.5 | 100% | 41 | 1.25 | 1.0 | 25% |
| 17 | 1.0 | 0.75 | 33.3% | 42 | 1.0 | 0.5 | 100% |
| 18 | 0.75 | 0.5 | 50% | 43 | 1.0 | 0.5 | 100% |
| 19 | 1.0 | 0.5 | 100% | 44 | 0.5 | 0 | $\infty$ |
| 20 | 0.75 | 0 | $\infty$ | 45 | 1.0 | 0.5 | 100% |
| 21 | 1.0 | 0 | $\infty$ | 46 | 1.0 | 0.5 | 100% |
| 22 | 0.75 | 0.5 | 50% | 47 | 0.75 | 0.5 | 50% |
| 23 | 1.0 | 0.5 | 100% | 48 | 0.25 | 0 | $\infty$ |
| 24 | 1.25 | 0.75 | 66.7% | 49 | 1.0 | 0.5 | 100% |
| 25 | 0.5 | 0 | $\infty$ | 50 | 1.0 | 0.5 | 100% |

### 2.6.4 Complete Results

Following the same setting as for the case study of one network instance in the Section 2.6.1, we randomly generate 50 instances, each with 20-node primary network and 30-node secondary network. For each instance, we randomly generate primary and secondary sessions, and compare the objective values obtained by the transparent coexistence paradigm and interference avoidance paradigm. Table 3.3 shows the results from 50 network instances. The fourth column shows the percentage improvement for transparent coexistence paradigm over interference avoidance paradigm. Note that some of the entries have $\infty$, indicating that the achievable session throughput (in DoFs) in the interference avoidance paradigm is $0$. Overall, we find that the achievable session throughput under the transparent coexistence paradigm is much higher than that under the interference avoidance paradigm.

## 2.7 Chapter Summary

This chapter explores the transparent coexistence paradigm for a multi-hop secondary network. This paradigm allows a secondary network to use the same spectrum simultaneously with the primary network as long as its activities are "transparent" (or "invisible") to the primary network. Such transparency is accomplished through a systematic interference cancelation (IC) by the secondary nodes without any impact on the primary network. The new technical challenges in a multi-hop network include channel/time slot scheduling, IC to/from primary network by the secondary network, and IC within the secondary network. We develop a rigorous mathematical modeling for a secondary multi-hop network in the transparent coexistence paradigm. As an application, we apply this model to study a throughput maximization problem with the objective of maximizing the minimum throughput among all secondary sessions. For the optimization problem, we develop an efficient polynomial time algorithm. Through simulation results, we show that the transparent coexistence paradigm offers significant improvement in spectrum access and throughput performance over the existing prevailing interference avoidance paradigm.

# Chapter 3

# Transparent Coexistence: A Distributed Algorithm

## 3.1 Introduction

A spectrum sharing paradigm is defined by how the secondary and the primary users achieve coexistence. In [22], Goldsmith *et al.* outlined three main paradigms, namely *interweave, underlay,* and *overlay.* Interweave is a simple but conservative approach that follows the traditional interference avoidance paradigm. Under interweave, a secondary network is allowed to access radio spectrum only when it is not in conflict with the primary network's user in time, frequency, or space [21, 26, 72]. On the other hand, overlay is considered an aggressive spectrum sharing paradigm as it encourages proactive cooperation between the primary and secondary networks in data forwarding [28, 31, 42, 61, 79, 83]. In terms of spectrum sharing efficiency and network performance, overlay represents the ultimate coexistence paradigm, although its actual adaptation and deployment may still be years away due to the need of significant change in primary users' behavior. In this research, we focus on the underlay paradigm, which is considered as a major step forward beyond the interweave paradigm while requiring minimal change on the primary network. The underlay refers to that secondary users may be active simultaneously with the primary users in

the same vicinity and in the same frequency, as long as the secondary user's interference to primary users are negligible (or below a given threshold).

Underlay coexistence paradigm has been explored in [23,33,85,86]. In [23], Gao *et al.* studied the transmission strategies for a MIMO secondary link with a primary link. They proposed a secondary transmission strategy consisting of environment learning, channel training, and data transmission. In [85], Zhang and Liang studied the transmission strategy for a single secondary MIMO link coexisting with multiple primary receivers with interference-power constraints. In [86], Zhang *et al.* studied the secondary-link beamforming pattern to achieve the coexistence of a single secondary link with multiple primary links. They aimed to maximize the secondary user's throughput while keeping the interference temperature at the primary receivers below a certain threshold. In [33], Kim and Giannakis studied the coexistence of multiple secondary links with one primary link. They proposed a distributed resource allocation algorithm to maximize the weighted sum rate of secondary links under a transmit power constraint at the secondary transmitters and an interference power constraint at the primary receiver. All these prior efforts were from information theoretic perspective. A common limitation of these prior efforts is that they are all limited to very simple network settings, e.g., several nodes or link pairs, all for single-hop communications.

In a recent study [75], we explored the underlay paradigm for a secondary multi-hop network under the name of *transparent coexistence (TC).* Under TC, there is no change on the primary network's behavior. It uses the spectrum as it wishes and is not concerned with the needs of the secondary network. On the other hand, the secondary network is allowed to access the spectrum in the *same* time, frequency, and location with the primary network, as long as its activities are "invisible" to the primary network. Such transparency is achieved by having the secondary network proactively cancel its interference to the primary network with powerful physical (PHY) layer techniques so that the primary nodes do not feel the presence of the secondary nodes. As a result, simultaneous activation of the secondary network along with the primary network is possible. In [75], we developed centralized mathematical models to characterize (i) *inter-network* interference cancelation (IC) relationships between two networks – secondary transmitters need to cancel their interference to the primary receivers while secondary receivers need to cancel the interference from

the primary transmitters; and (ii) *intra-network* IC – secondary nodes need to perform IC within their own network so that data can be transported successfully within the secondary network.

The results in [75] showed the concept of achieving TC for a multi-hop primary and secondary network through a centralized solution. But it is also desirable to have a distributed solution to achieve TC. The main contribution of this chapter is the development of a distributed scheduling algorithm for the secondary network to achieve TC with the primary network, while maximizing its own network throughput. For IC, we assume each secondary node is equipped with MIMO, while there is no requirement on the primary nodes. We employ a MIMO IC model that was developed in [59] to keep track of degree-of-freedoms (DoFs) allocation for transporting data streams (i.e., spatial multiplexing (SM)) and IC. It was shown in [59] that this IC model is efficient in DoF allocation while guaranteeing feasibility in the final solution. By feasibility, we mean there exists a feasible precoding and decoding vector for each data stream at the PHY layer. However, this model is centralized in nature and requires to maintain a global node ordering among the secondary nodes in the network, which is not possible in a distributed network environment. In this chapter, instead of maintaining a global node ordering, we only maintain two local sets at each node to keep track of the node's IC responsibilities. We show how to establish, maintain, and update these two local sets at each node in each iteration of our distributed algorithm. Our distributed algorithm increases the data stream on each active link iteratively based on local computation. Since the nodes in the two local sets of a node directly affect the node's IC responsibility, our algorithm attempts to switch nodes in the two sets if it can improve the IC structure. Although no explicit node ordering is maintained in our distributed algorithm, we prove that our distributed data structure at each node (with the use of two local sets) can be mapped to an explicit global node ordering for IC among all nodes in the network. From this global node ordering for IC among all nodes, we show there exist a set of feasible precoding vectors at each secondary transmitter and a feasible set of decoding vectors at each secondary receiver so that all data (in both primary and secondary networks) can be transported free of interference. Through numerical results, we show that the iterative distributed algorithm that we propose offers competitive performance when compared with an upper bound result from centralized optimization.

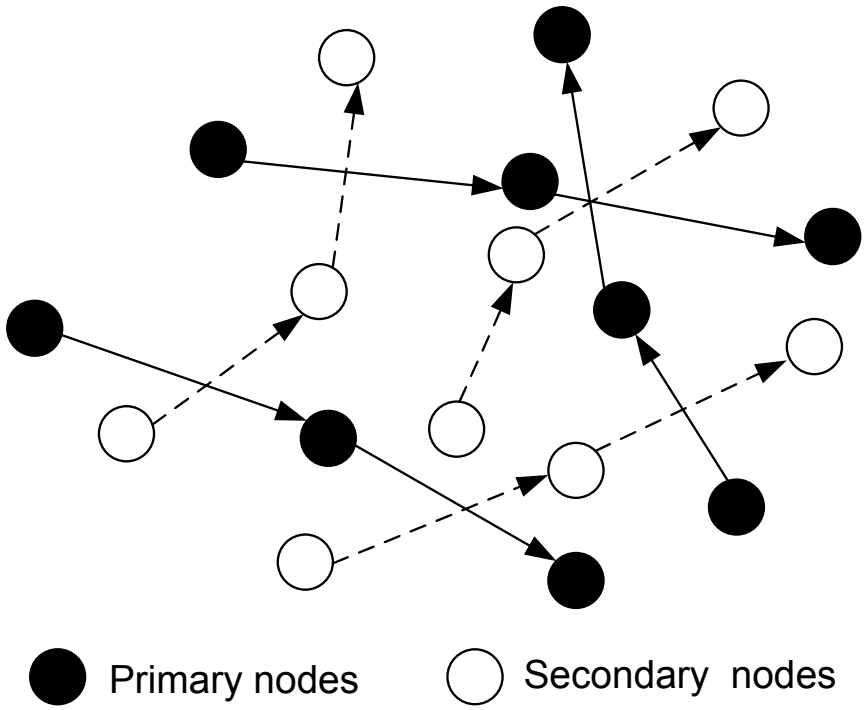



Figure 3.1: A multi-hop secondary network co-located in the same area as a multi-hop primary network.

The remainder of this chapter is organized as follows. In Section 3.2, we describe our problem. Section 6.3 presents the design of an iterative distributed algorithm to achieve TC for a secondary multi-hop network. In Section 3.4, we present a feasibility proof of our distributed algorithm at the PHY layer. In Section 3.5, we analyze the complexity our distributed algorithm. Section 6.4 presents numerical results and demonstrates the competitive performance of the proposed distributed algorithm. Section 6.5 concludes this chapter.

## 3.2 Problem Statement

In this chapter, we consider a multi-hop primary network (with a set of nodes $\mathcal{P}$) and a multi-hop secondary network (with a set of nodes $\mathcal{S}$) that are co-located in the same geographical area, as shown in Fig. 3.1. The primary network is assigned a certain spectrum band for its communi-

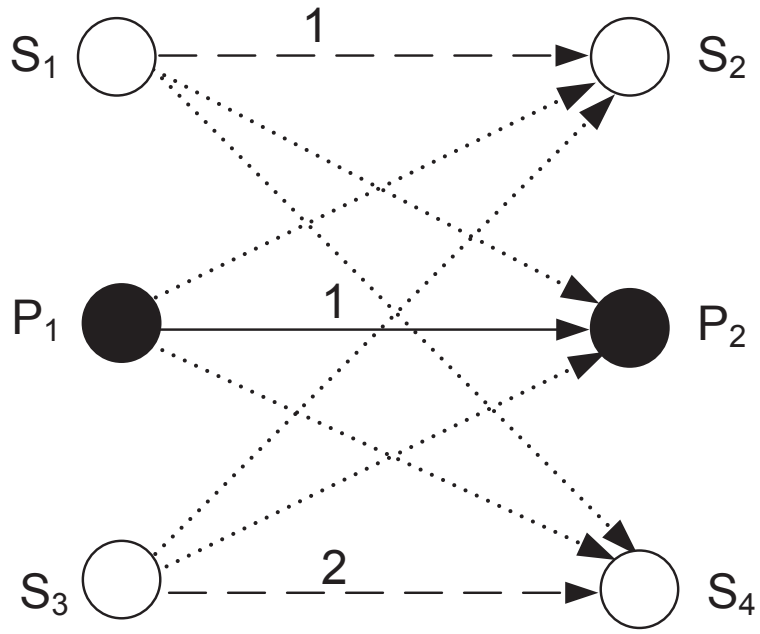



Figure 3.2: A simple example illustrating SM and IC. A solid line represents the primary link, a dashed line represents a secondary link, and a dotted line represents an interference.

cation. Suppose scheduling is done in the time domain, with $T$ time slots in a frame. For the primary network, it performs scheduling for transmission/reception without any consideration of the secondary network. A secondary node, however, is allowed to transmit in a time slot only if it is able to cancel its interference to its neighboring primary receivers. We assume the primary nodes are single antenna nodes. Suppose that there is a set of sessions $\tilde{\mathcal{F}}$ in the primary network $\mathcal{P}$. Each session has a source node and a destination node and traverses multi-hop relay nodes as needed. The route from a session's source node to its destination node is given a prior, which may be found by some standard routing protocols (e.g., AODV [45], DSR [32]). Denote $\tilde{\mathcal{L}}$ as a set of links in the network that are traversed by the active sessions in $\tilde{\mathcal{F}}$. Suppose the set of links $\tilde{\mathcal{L}}$ is operating under a feasible scheduling solution (for transmisson/reception) for the primary sessions $\tilde{\mathcal{F}}$, where interference at a primary receive node is avoided either through time slot or sufficient spatial separation. Since each primary node has only a single antenna, it can transmit at most one data stream to another node in a time slot.

For the secondary network, we assume each node is equipped with MIMO, which offers IC capability that is needed to achieve TC. We assume the number of antennas at a secondary node $i \in \mathcal{S}$ is $A_i$. For the secondary network $\mathcal{S}$, suppose that there is a set of sessions $\mathcal{F}$ in $\mathcal{S}$. Similar to a primary session, a secondary session has a source node, a destination node, and traverses multi-hop relay nodes as needed. The route from a secondary session's source node to its destination node is again given a priori. Denote $\mathcal{L}$ as the set of links that are traversed by any session in $\mathcal{F}$.

We use DoFs at a secondary node (no more than the number of antennas at the node) to represent its available resources. A DoF can be used for SM or IC. For SM, transmitting one data stream requires one DoF at the transmitter and one DoF at the receiver. In practice, the data rate carried in each data stream may vary with different channel conditions. For simplicity, we assume that the fixed modulation and coding scheme (MCS) is used for a link's data stream transmission, and one data stream in one time slot corresponds to one unit data rate. On the other hand, DoF consumption for IC depends on whether the IC is done at the transmitter or receiver. We use a simple example to illustrate this point. In Fig. 3.2, suppose $P_1$ and $P_2$ are a pair of primary transmit and receive nodes, while $S_1$ and $S_2$, $S_3$ and $S_4$ are two pairs of secondary transmit and receive nodes. Suppose that both the primary nodes $P_1$ and $P_2$ have one antenna, and the secondary nodes $S_1$, $S_2$, $S_3$, and $S_4$ are each equipped with $4$ antennas (4 DoFs). $P_1$ is transmitting $1$ data stream to $P_2$, $S_1$ is transmitting 1 data stream to $S_2$, and $S_3$ is transmitting 2 data stream to $S_4$. For the interference from $S_1$ to $S_4$, either transmitter $S_1$ or receiver $S_4$ can cancel this interference. If $S_1$ is to cancel this interference, then it will use 2 DoFs since $S_4$ is receiving 2 data streams; if $S_4$ is to cancel this interference, then it will use 1 DoF since $S_1$ is transmitting 1 data stream. Note the difference in DoF consumptions in IC by different nodes.

As described, to achieve TC, the secondary nodes have the sole responsibility to cancel interference to/from the primary nodes (i.e., inter-network interference) and interference within the secondary network (i.e., intra-network interference). In this example, for inter-network IC, secondary nodes $S_2$ and $S_4$ need to cancel the inference from primary transmit node $P_1$ with 1 DoF, respectively; the secondary transmit nodes $S_1$ and $S_3$ need to cancel their interference to primary receive node $P_2$ with 1 DoF, respectively. For intra-network IC, the interference from $S_1$ to $S_4$

needs to be cancelled, either by $S_1$ (with 2 DoF) or by $S_4$ (with 1 DoF) as discussed earlier; the interference from $S_3$ to $S_2$ needs to be cancelled, either by $S_3$ (with 1 DoF) or by $S_2$ (with 2 DoFs).

To successful perform inter- and intra-network IC, it is crucial for the secondary nodes to have accurate channel state information (CSI). A practical problem to address is: how can a secondary node obtain the CSI between itself and its neighboring primary nodes while remaining transparent to the primary nodes? There are many schemes that have been proposed to address this issue (see, e.g., [49, 62, 74, 75, 77, 84]). We omit their discussions here to conserve space. But the point here is that there exist schemes that we can use to obtain the necessary CSI for the secondary nodes to perform inter- and intra-network IC.

In our design of distributed algorithm for the secondary nodes to achieve TC, we consider a throughput maximization problem, with the objective of maximizing the minimum achievable session rate (in terms of data streams) among all secondary sessions. We choose this objective since it focuses on the worst case (minimum) achievable secondary session throughput, which ensures fairness across all secondary sessions.

## 3.3 A Distributed Algorithm

We propose a distributed scheduling algorithm to the throughput maximization problem while meeting all IC requirements for the secondary nodes. As described in our network setting, the set of sessions $\tilde{\mathcal{F}}$ in the primary network are transmitting under a given feasible scheduling solution. To have the secondary sessions operate in the same set of time slots (to achieve TC), we employ MIMO at the secondary nodes for IC. The algorithm that we propose is an iterative greedy algorithm. We consider one link (from the set of links $\mathcal{L}$) at a time and try to increase the data streams on this link by 1 in this iteration. This increment is successful only if the transmitter, receiver and neighboring nodes of this link have enough remaining DoFs to cancel this new interference on neighboring primary and secondary nodes.

As discussed earlier, an interference can be canceled either by a secondary transmit or receive

node. For efficient and feasible IC, a global node ordering scheme proposed in [59] would be useful. But such a global node ordering scheme is centralized in nature. Nevertheless, it gives us some hints in our design of distributed algorithm.

We propose to maintain two local sets at each node to keep track of the IC responsibility between this node and neighboring nodes. For example, at each secondary node $i \in \mathcal{S}$, we maintain one local set $\mathcal{B}_i(t)$ to store $i$'s neighboring nodes that require node $i$ to use its DoFs for IC and the other local set $\mathcal{Y}_i(t)$ to store $i$'s neighboring nodes that use their own DoFs for canceling interference to/from node $i$ (see Fig. 3.3). Note that there is no explicit node ordering among the nodes in sets $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$. By maintaining these two sets (with $\mathcal{B}_i(t)$ before node $i$ and $\mathcal{Y}_i(t)$ after node $i$), we have achieved the desired efficiency in IC locally at node $i$. We will discuss the feasibility issue in Section 3.4.

The use of two local sets $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$ at each secondary node $i$ is centerpiece in our design of distributed scheduling algorithm to achieve TC. In our algorithm, we will exploit these two sets at each node to its fullest extent to achieve IC at the secondary nodes while meeting the resource constraints (limited DoFs at each node). In particular, when we find that a data stream cannot be further increased on a bottleneck link, we will consider moving some nodes from one local set into the other set so that the DoFs at a node can be re-allocated. This step is called adjusting IC responsibility in our algorithm (Step 3) and is a critical component to maximize the performance of our algorithm. At any iteration when this IC responsibility adjustment is not successful (and thus the number of data streams on the associated link cannot be further increased) for all time slots in a frame, our algorithm terminates.

### 3.3.1 State Information at Secondary Nodes

The state information that needs to be maintained at a secondary node (say $i$) is shown in Table 3.1.

**Local sets $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$:** For each interference involving node $i$, it can be canceled by either node $i$ or the other node involved in this interference. To explicitly distinguish who is responsible

Table 3.1: State information at each node $i$

| Symbol | Definition |
|---|---|
| $s_i(t)$ | The status of node $i$ in time slot $t$. $s_i(t)$ can be either Tx, Rx or Idle. |
| $\mathcal{B}_i(t)$ | The set of nodes that node $i$ allocates DoFs for IC to/from them in time slot $t$. |
| $\mathcal{Y}_i(t)$ | The set of nodes that allocate their own DoFs for IC to/from node $i$ in time slot $t$. |
| $\lambda_i^{\mathrm{SM}}(t)$ | The number of DoFs that node $i$ has allocated for SM in time slot $t$. |
| $\lambda_i^{\mathrm{IC}}(t)$ | The number of DoFs that node $i$ has allocated for IC in time slot $t$. |
| $\lambda_i^{\mathrm{RM}}(t)$ | The number of remaining DoFs at node $i \in \mathcal{S}$ in time slot $t$, i.e., $\lambda_i^{\mathrm{RM}}(t) = A_i - \lambda_i^{\mathrm{SM}}(t) - \lambda_i^{\mathrm{IC}}(t)$. |
| $\tilde{\alpha}_i(t)$ | The total number of data streams from node $i$'s neighboring primary transmitters in time slot $t$. |
| $\tilde{\beta}_i(t)$ | The total number of data streams received by node $i$'s neighboring primary receivers in time slot $t$. |
| $\alpha_i(t)$ | The total number of data streams from node $i$'s neighboring secondary transmitters in time slot $t$. |
| $\beta_i(t)$ | The total number of data stream received by node $i$'s neighboring secondary receivers in time slot $t$. |
| $z_{i,j}(t)$ | The number of data streams from transmit node $i$ to receive node $j$. |

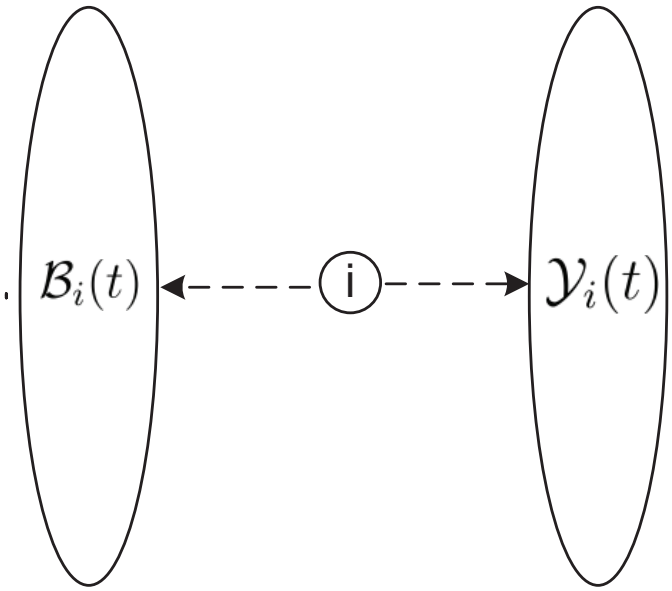



Figure 3.3: Maintaining two local sets at node $i$ to distinguish IC responsibility between node $i$ and its neighboring nodes.

for IC for each interference, we maintain two local sets $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$ at each node $i$, as shown in Figure 3.3. We denote $\mathcal{B}_i(t)$ as the set of secondary nodes that node $i$ ($i \in \mathcal{S}$) allocates DoFs to cancel interference to/from them, and denote $\mathcal{Y}_i(t)$ as the set of secondary nodes that allocate their DoFs to cancel interference to/from $i$. At the beginning of our algorithm, we initialize $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$ as empty sets, i.e., $\mathcal{B}_i(t) = \emptyset$ and $\mathcal{Y}_i(t) = \emptyset$ for $i \in \mathcal{S}$.

**Accounting of DoF resource:** In Table 3.1, $z_{i,j}(t)$ represents the number of data stream transmitted from node $i$ to node $j$. $\lambda_i^{\mathrm{SM}}(t)$ and $\lambda_i^{\mathrm{IC}}(t)$ represents the number of DoFs allocated for SM and IC at secondary node $i$ in time slot $t$, respectively. $\lambda_i^{\mathrm{RM}}(t)$ represents the number of remaining DoFs at a node $i$ in time slot $t$. At the beginning of our algorithm, the status of each node $i \in \mathcal{S}$ is set to Idle, i.e., $s_i(t)$ =Idle for $t = 1, 2, \cdots, T$. Then, the initial DoF allocation for SM and IC at each node is 0. We have $\lambda_i^{\mathrm{SM}}(t) = \lambda_i^{\mathrm{IC}}(t) = 0$, $\lambda_i^{\mathrm{RM}}(t) = A_i$ and $z_{i,j}(t) = 0$ for $i, j \in \mathcal{S}, t = 1, 2, \cdots, T$ in the initialization stage. $\tilde{\alpha}_i(t)$ and $\tilde{\beta}_i(t)$ are constants and are calculated based on active sessions in the primary network. These can be derived by the secondary nodes through monitoring/sensing of the neighboring primary nodes's activities. On the other hand, the

initial values for $\alpha_i(t)$ and $\beta_i(t)$ are 0.

For these state information, except that $\tilde{\alpha}_i(t)$ and $\tilde{\beta}_i(t)$ are constants, the values for $s_i(t)$, $\mathcal{B}_i(t)$, $\mathcal{Y}_i(t)$, $\lambda_i^{\mathrm{SM}}(t)$, $\lambda_i^{\mathrm{IC}}(t)$, $\lambda_i^{\mathrm{RM}}(t)$, $z_{i,j}(t)$, $\alpha_i(t)$ and $\beta_i(t)$ are variables and will be updated during each iteration of the algorithm.

### 3.3.2 Step 1: Link Selection

To make a rate increment of each session by 1 DoF is equivalent to increasing the DoF on each active link by 1 DoF if each active link is traversed by 1 session. In the general case when an active link is traversed by multiple sessions, we need to increase the DoFs on this active link by multiple times, each for one session. In our distributed algorithm, we choose an active link for increment during an iteration. If a link is traversed by multiple sessions, then it is necessary to represent the link multiple times so that each session traversing this link is to be considered for data stream increment. Suppose there are $k$ sessions traversing a link $l \in \mathcal{L}$. Then we represent link $l$ by $k$ *logical* links. We want to set a round robin for these logical links for rate increment so that each logical link is considered once in each cycle.

To do this, we employ the so-called distributed ranking algorithm by Zaks [81]. This algorithm was designed to solve the problem of sorting and ranking $n$ processors in a distributed system. The input is an initial value unique for each processor. The output is a ranking of all $n$ processors. To apply the distributed ranking algorithm, we assign an initial value for each logical link. Each initial value is generated randomly and guaranteed to be unique (under a reasonably good random number generator). We let the transmitter of each logical link to maintain the logical link's rank. After a logical link obtains its rank, it will know precisely when it will be considered for data stream increment.

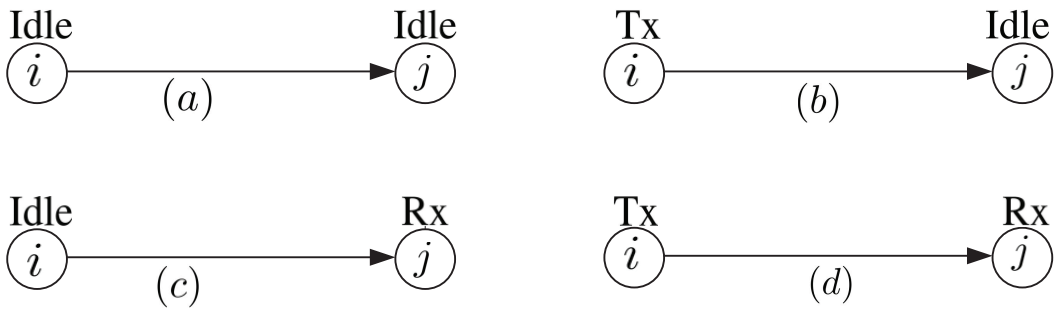



Figure 3.4: Four cases of link status.

### 3.3.3 Step 2: Data Stream Increment

After we identify a logical link (in Step 1), our algorithm will try to increase one data stream on the selected link, while satisfying IC constraints and transparency to the primary network.[1] We first present the necessary conditions under which one more data stream can be added on the link in a time slot. Then we describe how to update state information on the nodes that are involved in this increment.

**Sufficient Conditions for Data Stream Increment.** We now discuss when the number of data streams on a chosen link can be increased by 1 in a given time slot. Suppose link $(i, j)$ is the link. Then both nodes $i$ and $j$ first check their current status ("Tx", "Rx", or "Idle"). Some cases can be clearly ruled out for consideration, i.e., $s_i(t) =$ Rx or $s_j(t) =$ Tx. In these cases, link $(i, j)$ cannot be considered for data stream increment in time slot $t$ and we move to the next time slot $(t + 1)$ immediately. When link $(i, j)$ is suitable for data stream increment, there are four possible statuses as shown in Figure 3.4. The sufficient conditions for data stream increment on link $(i, j)$ are as follows.

*Case (a):* $s_i(t) =$ Idle and $s_j(t) =$ Idle.

- $s_i(t) =$ Idle: Since node $i$ is idle, the local sets $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$ are empty. We need to establish the sets $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$ (see Figure 3.3) to decide the IC relationships between node $i$ and its neighboring secondary receive nodes that will be interfered by $i$. We can put

[1] We drop the fine distinction between "link" and "logical link" when there is no confusion.

all these neighboring receive nodes in time slot $t$ either in $\mathcal{B}_i(t)$ or $\mathcal{Y}_i(t)$.

- If all neighboring receive nodes are put into $\mathcal{Y}_i(t)$, then the interference from node $i$ to them will be canceled by these receive nodes. The following two conditions must be satisfied: (i) the total number of DoFs at node $i$ should be greater than the total number of data streams received by its neighboring primary receivers, i.e., $A_i > \tilde{\beta}_i(t)$, (ii) all secondary receivers that are in $\mathcal{Y}_i(t)$ must have at least one remaining DoFs to cancel one more data stream interference from node $i$.
- If all neighboring receive nodes are put into $\mathcal{B}_i(t)$, node $i$ needs to cancel its interference to all these neighboring receive nodes. The following condition must be satisfied: the total number of DoFs at node $i$ is more than the sum of data streams received by both neighboring primary and secondary receivers, i.e., $A_i > \tilde{\beta}_i(t) + \beta_i(t)$.

- $s_j(t) =$ Idle: Similar to node $i$, we put node $j$'s neighboring transmit nodes in either $\mathcal{B}_j(t)$ or $\mathcal{Y}_j(t)$.
  - If all neighboring transmit nodes are put into $\mathcal{Y}_j(t)$, these transmit nodes should cancel their interference to node $j$. Then the following two conditions must be satisfied: (i) the total number of DoFs at node $j$ should be greater than the total number of data streams transmitted by its neighboring primary transmitters, i.e., $A_j > \tilde{\alpha}_j(t)$, (ii) all secondary transmitters that are in $\mathcal{Y}_j(t)$ must have at least one remaining DoFs to cancel its interference to node $j$.
  - If all neighboring transmit nodes are put into $\mathcal{B}_j(t)$, node $j$ should cancel interference from all these transmit nodes. The following condition must be satisfied: the total number of DoFs at node $j$ is more than the sum of data streams transmitted by both neighboring primary and secondary transmitters, i.e., $A_j > \tilde{\alpha}_j(t) + \alpha_j(t)$.

If the conditions for $s_i(t) =$ Idle and $s_j(t) =$ Idle are both satisfied, we proceed with this increment and update state information at nodes $i, j$ and their neighboring nodes according to Figure 3.5 and Figure 3.6.

**State update at idle node $i$ and neighboring receiver $k$**

1. If $s_i(t) =$ Idle:
2. Update $s_i(t) =$ Tx; $\lambda_i^{\mathrm{SM}}(t) \leftarrow \lambda_i^{\mathrm{SM}}(t) + 1$;
   $\lambda_i^{\mathrm{RM}}(t) \leftarrow \lambda_i^{\mathrm{RM}}(t) - 1$; $z_{i,j}(t) \leftarrow z_{i,j}(t) + 1$.
3. If $\mathcal{Y}_i(t) \leftarrow \{$Neighboring active secondary receivers.$\}$
4. Update $\lambda_i^{\mathrm{IC}}(t) \leftarrow \tilde{\beta}_i(t)$; $\lambda_i^{\mathrm{RM}}(t) \leftarrow \lambda_i^{\mathrm{RM}}(t) - \lambda_i^{\mathrm{IC}}(t)$;
5. For each receive node $k \in \mathcal{Y}_i(t)$:
6. $\mathcal{B}_k(t) \leftarrow \mathcal{B}_k(t) \cup \{i\}$; $\lambda_k^{\mathrm{IC}}(t) \leftarrow \lambda_k^{\mathrm{IC}}(t) + 1$; $\lambda_k^{\mathrm{RM}}(t) \leftarrow \lambda_k^{\mathrm{RM}}(t) - 1$.
7. Else if $\mathcal{B}_i(t) \leftarrow \{$Neighboring active secondary receivers.$\}$
8. Update $\lambda_i^{\mathrm{IC}}(t) \leftarrow \tilde{\beta}_i(t) + \beta_i(t)$; $\lambda_i^{\mathrm{RM}}(t) \leftarrow \lambda_i^{\mathrm{RM}}(t) - \lambda_i^{\mathrm{IC}}(t)$.
9. For each node $k \in \mathcal{B}_i(t)$: $\mathcal{Y}_k(t) = \mathcal{Y}_k(t) \cup \{i\}$.

Figure 3.5: Pseudocode to update state information when $s_i(t) =$ Idle.

*Case (b):* $s_i(t) =$ Tx and $s_j(t) =$ Idle.

- $s_i(t) =$ Tx : In this case, the following conditions must be satisfied if node $i$ wants to increase one more data stream on link $(i, j)$: (i) node $i$ has at least one remaining DoF for SM, i.e., $\lambda_i^{\mathrm{RM}}(t) \geq 1$; (ii) each receive node $k \in \mathcal{Y}_i(t)$ has at least one remaining DoF to cancel the new interference from node $i$.

- $s_j(t) =$ Idle : This case has been discussed in *Case (a)*.

If the conditions for $s_i(t) =$ Tx and $s_j(t) =$ Idle are both satisfied, we proceed with this increment and update state information at nodes $i, j$ and their neighboring nodes according to Figure 3.6 and Figure 3.7.

*Case (c):* $s_i(t) =$ Idle and $s_j(t) =$ Rx.

- $s_i(t) =$ Idle: This case has been discussed in Case (a).

- $s_j(t) =$ Rx: In this case, the following condition must be satisfied if node $j$ wants to increase one more data stream on link $(i, j)$: (i) node $j$ has at least one remaining DoF for SM, i.e.,

**State update at idle node $j$ and neighboring transmitter $k$**

1. If $s_j(t) =$ Idle:
2. Update $s_j(t) =$ Rx; $\lambda_j^{\mathrm{SM}}(t) \leftarrow \lambda_j^{\mathrm{SM}}(t) + 1$;
   $\lambda_j^{\mathrm{RM}}(t) \leftarrow \lambda_j^{\mathrm{RM}}(t) - 1$; $z_{i,j}(t) \leftarrow z_{i,j}(t) + 1$.
3. If $\mathcal{Y}_j(t) \leftarrow \{$Neighboring active secondary transmitters.$\}$
4. Update $\lambda_j^{\mathrm{IC}}(t) \leftarrow \tilde{\alpha}_j(t)$; $\lambda_j^{\mathrm{RM}}(t) \leftarrow \lambda_j^{\mathrm{RM}}(t) - \lambda_j^{\mathrm{IC}}(t)$
5. For each transmit node $k \in \mathcal{Y}_j(t)$:
6. $\mathcal{B}_k(t) \leftarrow \mathcal{B}_k(t) \cup \{j\}$; $\lambda_k^{\mathrm{IC}}(t) = \lambda_k^{\mathrm{IC}}(t) + 1$; $\lambda_k^{\mathrm{RM}}(t) = \lambda_k^{\mathrm{RM}}(t) - 1$.
7. Else if $\mathcal{B}_j(t) \leftarrow \{$Neighboring active secondary transmitters.$\}$
8. Update $\lambda_j^{\mathrm{IC}}(t) \leftarrow \tilde{\alpha}_j(t) + \alpha_j(t)$; $\lambda_j^{\mathrm{RM}}(t) \leftarrow \lambda_j^{\mathrm{RM}}(t) - \lambda_j^{\mathrm{IC}}(t)$.
9. For each $k \in \mathcal{B}_j(t)$: $\mathcal{Y}_k(t) \leftarrow \mathcal{Y}_k(t) \cup \{j\}$.

Figure 3.6: Pseudocode to update state information when $s_j(t) =$ Idle.

**State update at transmit node $i$ and neighboring receiver $k$**

1. If $s_i(t) =$ Tx:
2. Update $\lambda_i^{\mathrm{SM}}(t) \leftarrow \lambda_i^{\mathrm{SM}}(t) + 1$; $\lambda_i^{\mathrm{RM}}(t) \leftarrow \lambda_i^{\mathrm{RM}}(t) - 1$; $z_{i,j}(t) = z_{i,j}(t) + 1$.
3. For each receive node $k \in \mathcal{Y}_i(t)$:
4. Update $\lambda_k^{\mathrm{IC}}(t) \leftarrow \lambda_k^{\mathrm{IC}}(t) + 1$; $\lambda_k^{\mathrm{RM}}(t) \leftarrow \lambda_k^{\mathrm{RM}}(t) - 1$.

Figure 3.7: Pseudocode to update state information when $s_i(t) =$ Tx.

$\lambda_j^{\mathrm{RM}}(t) \geq 1$; (ii) each transmit node $k \in \mathcal{Y}_j(t)$ has at least one remaining DoF to cancel its interference to node $j$.

If the conditions for $s_i(t) =$ Idle and $s_j(t) =$ Rx are both satisfied, we proceed with this increment and update state information at nodes $i, j$ and their neighboring nodes according to Figure 3.5 and Figure 3.8.

*Case (d):* $s_i(t) =$ Tx and $s_j(t) =$ Rx. The case for $s_i(t) =$ Tx has been discussed in *Case (b)* and $s_j(t) =$ Rx has been discussed in *Case (c)*. If the conditions for $s_i(t) =$ Tx and $s_j(t) =$ Rx are both satisfied, we proceed with this increment and update state information at nodes $i, j$ and their neighboring nodes according to Figure 3.7 and Figure 3.8.

**State update at receive node $j$ and neighboring transmitter $k$**

1. If $s_j(t) = \text{Rx}$:
2. Update $\lambda_j^{\text{SM}}(t) \leftarrow \lambda_j^{\text{SM}}(t) + 1$; $\lambda_j^{\text{RM}}(t) \leftarrow \lambda_j^{\text{RM}}(t) - 1$; $z_{i,j}(t) = z_{i,j}(t) + 1$.
3. For each transmit node $k \in \mathcal{Y}_j(t)$:
4. Update $\lambda_k^{\text{IC}}(t) \leftarrow \lambda_k^{\text{IC}}(t) + 1$; $\lambda_k^{\text{RM}}(t) \leftarrow \lambda_k^{\text{RM}}(t) - 1$.

Figure 3.8: Pseudocode to update state information when $s_j(t) = \text{Rx}$.

Recall that there are $T$ time slots in a time frame. Node activities (both primary and secondary) and interference patterns in each time slot are different. If the data stream increment operation described above fails in the first time slot, we try it again in the second time slot and so forth, until a data stream increment is successful in a time slot or fails after all $T$ time slots.

### 3.3.4 Step 3: Adjusting Node's IC Responsibility

If the sufficient conditions at either node $i$ or node $j$ cannot be satisfied, we move on to this step. The only reason why link $(i, j)$ fails to increase one data stream in step 2 is the lack of DoF resources at some nodes (bottleneck nodes), i.e., node $i, j$ or nodes in $\mathcal{Y}_i(t)$ and $\mathcal{Y}_j(t)$. Since a node's local sets $\mathcal{B}$ and $\mathcal{Y}$ directly affects its DoF consumption for IC, we will try to swap some nodes between the sets $\mathcal{B}$ and $\mathcal{Y}$, and thus change their IC responsibilities. For example, if node $k$ is short on DoFs, we can move some node $m \in \mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$, thereby transferring the IC responsibility from $k$ to $m$. Through this change, some new DoF resources for the bottleneck node $k$ become available, possibly allowing a new data stream increment to be made on the link under consideration.

The main idea of this step is as follows. For each time slot $t$, we identify the set of bottleneck nodes (denoted as $\mathcal{D}_{(i,j)}(t)$), which do not have enough remaining DoF resources should one more data stream is added onto link $(i, j)$. For each node $k \in \mathcal{D}_{(i,j)}(t)$, we adjust node $k$'s IC responsibility by moving some other nodes in $\mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$. To ensure feasibility, only a subset of nodes (denoted as $\bar{\mathcal{B}}_k(t)$), $\bar{\mathcal{B}}_k(t) \subseteq \mathcal{B}_k(t)$, is eligible for moving from $\mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$. After

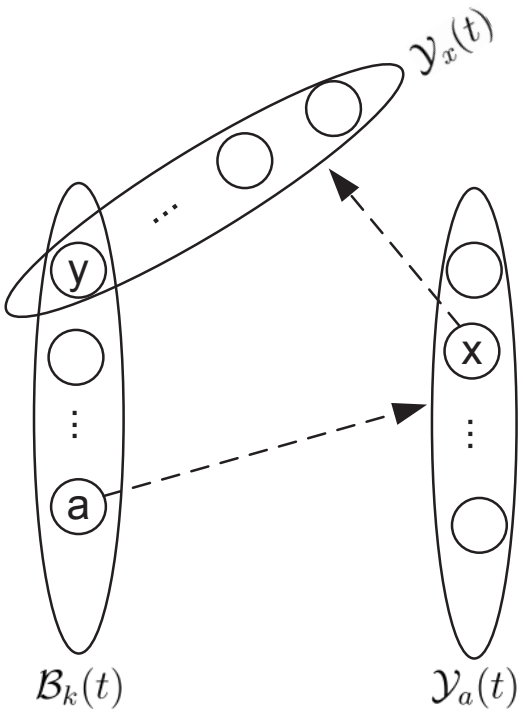
identifying $\bar{\mathcal{B}}_k(t)$ for $k$, we consider nodes in $\bar{\mathcal{B}}_k(t)$ in the order of non-increasing remaining DoFs, i.e., starting with the one that has the maximum remaining DoF (denoted as node $a$) if it is moved to $\mathcal{Y}_k(t)$. If this movement is infeasible, then our attempted adjustment fails in this time slot and we move on to the next time slot. Otherwise, we move $a$ from $\mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$ and update their state information. After this movement, if a new data stream can be added on link $(i, j)$, we are done. Otherwise, we continue moving the next node in $\bar{\mathcal{B}}_k(t)$ that has the maximum remaining DoF (denoted as node $b$) to $\mathcal{Y}_k(t)$ following the same process. This step terminates upon a new data stream can be successfully added on link $(i, j)$ or all nodes in $\mathcal{D}_{(i,j)}(t)$ are considered for all time slots in a frame. In the rest of this section, we give more details for this idea.

**Finding bottleneck nodes $\mathcal{D}_{(i,j)}(t)$:** $\mathcal{D}_{(i,j)}(t)$ can be easily found by identifying those nodes that would need more DoFs than their remaining DoFs should one more data stream were added on link $(i, j)$.

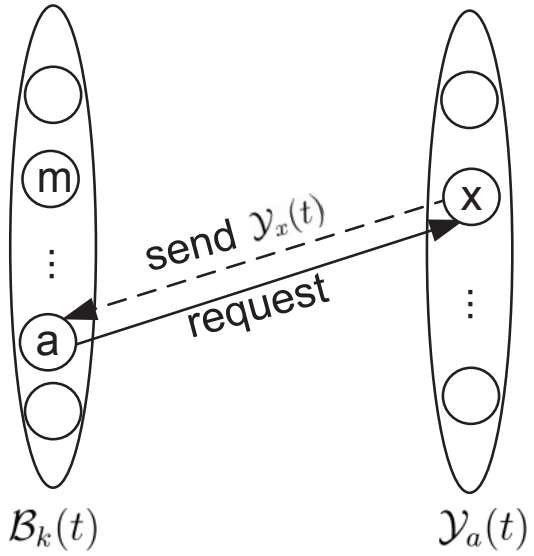
**Node sequence in $\mathcal{D}_{(i,j)}(t)$:** To consider nodes one at a time in $\mathcal{D}_{(i,j)}(t)$ in a distributed environment, we could use a token to pass along from one node to the next so that at any time, only one node is considered for adjustment. There is no preference on which node to start but for the rest of the discussion, we assume that we start with node $i$, then $j$, before the other nodes in $\mathcal{D}_{(i,j)}(t)$. Note that a token is passed to the next node in $\mathcal{D}_{(i,j)}(t)$ only if the adjustment in the previous node is successful. Otherwise, the algorithm moves on to the next time slot in the frame.

**Finding eligible subset nodes for swapping:** Suppose the token is now passed onto node $k \in \mathcal{D}_{(i,j)}(t)$. To adjust node $k$'s IC responsibility, we want to move one or more nodes in $\mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$, thus relieving node $k$'s IC responsibility for these nodes. But for feasibility, not every node in $\mathcal{B}_k(t)$ is eligible for swapping. Now we discuss how to identify a subset of nodes $\bar{\mathcal{B}}_k(t)$ that are eligible to be moved to $\mathcal{Y}_k(t)$. By "eligible", we mean that when we move the subset of nodes from $\mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$, the IC responsibilities for all other nodes in $\bar{\mathcal{B}}_k(t)$ and $\mathcal{Y}_k(t)$ are not affected. We propose a sufficient condition to check whether or not a node is an eligible node as follows.

Suppose node $k$ is a transmitter. Then it can consider those receive nodes in $\mathcal{B}_k(t)$ for moving to $\mathcal{Y}_k(t)$. We denote $c \leftarrow b$ as node $b$ cancels interference to or from $c$. For a receive node $a \in \mathcal{B}_k(t)$,




(a) An illustration of determining the eligibility of receive node $a \in \mathcal{B}_k(t)$.




(b) Steps involved in determining the eligibility of receive node $a \in \mathcal{B}_k(t)$.

Figure 3.9: Determining the eligibility of receive node $a \in \mathcal{B}_k(t)$ when node $k$ is a transmit node.

it can be moved to $\mathcal{Y}_k(t)$ if the following conditions are satisfied. For each transmit node $x \in \mathcal{Y}_a(t)$ that needs to do IC to $a$ (i.e., $a \leftarrow x$), there cannot exist a receive node $y \in \mathcal{Y}_x(t)$ that $y$ handles IC from $x$ (i.e., $x \leftarrow y$), and $k$ handles IC to $y$ (i.e., $y \leftarrow k$) (see Figure 3.9 (a)). That is, there does not exist a receive node $y$, such that the following IC relationship holds: $a \leftarrow x \leftarrow y \leftarrow k$. If this condition is satisfied and $a$'s remaining DoFs is at least one after moving to $\mathcal{Y}_k(t)$, $a$ is an eligible node; otherwise, $a$ is not. In Section 3.4, we will show that this condition can guarantee IC feasibility among all nodes.

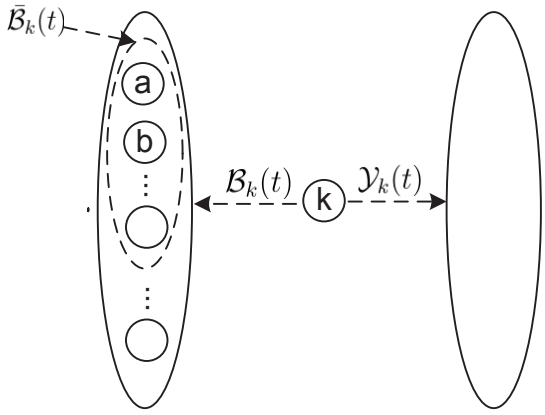
To do this check, we have node $a$ send a request for state information to those transmit nodes in $\mathcal{Y}_a(t)$. Upon receiving this request, each transmit node $x \in \mathcal{Y}_a(t)$ will send its state information $\mathcal{Y}_x(t)$ to node $a$ (see Figure 3.9(b)). Upon receiving this state information, node $a$ can check whether some receive nodes in $\mathcal{Y}_x(t)$ are also in $\mathcal{B}_k(t)$. If none of these receive nodes are in $\mathcal{B}_k(t)$ *and* $a$'s remaining DoFs is at least one after moving to $\mathcal{Y}_k(t)$, then $a$ is eligible. Otherwise, $a$ is not eligible.

The above discussion is for the case when node $k$ is a transmit node. The case when node $k$ is a receive or idle node can be handled in a similar manner.

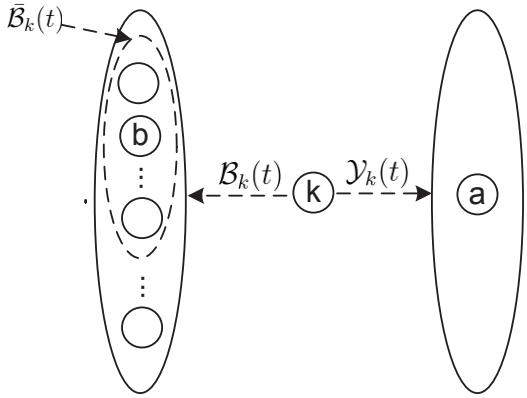
**Moving node(s) in $\bar{\mathcal{B}}_k(t)$ to $\mathcal{Y}_k(t)$:** Assume node $a \in \bar{\mathcal{B}}_k(t)$ has the maximum remaining DoFs after moving to $\mathcal{Y}_k(t)$. If $\bar{\mathcal{B}}_k(t) = \emptyset$, there is no eligible node and we move to next time slot immediately. Otherwise, at node $k$, we move $a$ from $\bar{\mathcal{B}}_k(t)$ to $\mathcal{Y}_k(t)$ while at node $a$, we move $k$ from $\mathcal{Y}_a(t)$ to $\mathcal{B}_a(t)$, and update their state information as follows.

- Case $s_k(t) =$ Tx or $s_k(t) =$ Rx: In this case, $k$ only needs to release one DoF. Since at node $k$, we move $a$ from $\mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$ while at node $a$, we move $k$ from $\mathcal{Y}_a(t)$ to $\mathcal{B}_a(t)$, then at least one DoF can be released from $k$. The node $k$ updates the state information based on Figure 3.11, and the node $a$ updates its state information based on Figure 3.12.

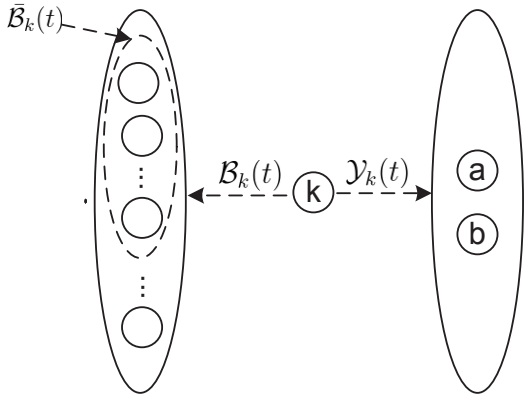
- Case $s_k(t) =$ Idle: Recall that for the bottleneck node $k$ in $\mathcal{D}_{(i,j)}(t)$, it might be $i, j$, or node in $\mathcal{Y}_i(t)$ or in $\mathcal{Y}_j(t)$. Since $\mathcal{Y}_i(t)$ represents the set of $i$'s neighboring receive nodes that should allocate DoFs to cancel interference from node $i$, and $\mathcal{Y}_j(t)$ represents the set of




(a) Before movement




(b) Move a to $\mathcal{Y}_k(t)$




(c) Move b to $\mathcal{Y}_k(t)$

Figure 3.10: An illustration of movement process when node $k$ is an idle node.

$j$'s neighboring transmit nodes that should allocate DoFs to cancel their interference to node $j$, then all nodes in $\mathcal{Y}_i(t)$ and $\mathcal{Y}_j(t)$ are active nodes. Therefore, node $k$ can only represent node $i$ or node $j$. Let's consider the case when node $k$ is node $i$. The case when node $k$ is node $j$ is similar. Recall that when $s_k(t) =$ Idle, both $\mathcal{B}_k(t)$ and $\mathcal{Y}_k(t)$ are empty. We establish $\mathcal{B}_k(t)$ and $\mathcal{Y}_k(t)$ by putting all neighboring active nodes in either $\mathcal{B}_k(t)$ or $\mathcal{Y}_k(t)$. Clearly, putting all neighboring receive nodes in $\mathcal{Y}_k(t)$ will add additional IC burden on all these nodes in $\mathcal{Y}_k(t)$ and may require adjusting each node's IC responsibility in $\mathcal{Y}_k(t)$. On the other hand, putting all neighboring receive nodes to $\mathcal{B}_k(t)$ will not have this issue as the IC responsibility on those nodes in $\mathcal{B}_k(t)$ are not affected and we only need to focus on adjusting node $k$'s responsibility with one node in $\mathcal{B}_k(t)$. We adopt the latter approach and put all neighboring receive nodes in $\mathcal{B}_k(t)$ and set $\mathcal{Y}_k(t) = \emptyset$ (see Figure 3.10 (a)). For each node $p \in \mathcal{B}_k(t)$, node $k$ is added to $\mathcal{Y}_p(t)$. Therefore, $\lambda_k^{\mathrm{IC}}(t) = \sum_{n\in\mathcal{B}_k(t)}^{s_n(t)=\mathrm{Rx}} \lambda_n^{\mathrm{SM}}(t) + \tilde{\beta}_k(t)$, where $\tilde{\beta}_k(t)$ is the total number of data streams received by node $i$'s neighboring primary receivers, and $\lambda_k^{\mathrm{RM}}(t) = A_k - \lambda_k^{\mathrm{IC}}(t)$. We start to put node $a$ (the node in $\bar{\mathcal{B}}_k(t)$ that has the maximum remaining DoFs after movement) (see Figure 3.10 (b)) into $\mathcal{Y}_k(t)$ . Both nodes $k$ and $a$'s state information is updated based on Figures 3.11 and 3.12, respectively. For the new sets, if a new data stream can be added on link $(i, j)$, we are done. Otherwise, we continue to move another node $b \in \bar{\mathcal{B}}_k(t)$ that has the maximum remaining DoFs after movement following the same process (see Figure 3.10(c)). The process terminates if node $k$ has at least one remaining DoF or $\bar{\mathcal{B}}_k(t) = \emptyset$. For the latter case, the adjustment fails and we move on to the next time slot.

## 3.4 Physical Layer Feasibility

In our design of distributed algorithm, for each node $k$, we put its neighboring nodes in two sets: $\mathcal{B}_k(t)$ and $\mathcal{Y}_k(t)$. For the set of nodes in $\mathcal{B}_k(t)$, node $k$ is responsible to cancel its interference to them if $k$ is a transmit node or cancel the interference from them if $k$ is a receive node. For the ease of understanding, we can consider the set of nodes in $\mathcal{B}_k(t)$ being positioned *before* node $k$ while

**State Information Update at node $k$**

1. Update $\mathcal{B}_k(t) \leftarrow \mathcal{B}_k(t) - \{a\}$, $\mathcal{Y}_k(t) \leftarrow \mathcal{Y}_k(t) \cup \{a\}$;
2. If $s_k(t) =$ Tx:
3. Update $\lambda_k^{\mathrm{IC}}(t) \leftarrow \lambda_k^{\mathrm{IC}}(t) - (\lambda_a^{\mathrm{SM}}(t) - z_{k,a}(t))$; $\lambda_k^{\mathrm{RM}}(t) \leftarrow \lambda_k^{\mathrm{RM}}(t) + (\lambda_a^{\mathrm{SM}}(t) - z_{k,a}(t))$;
4. Else if $s_k(t) =$ Rx:
5. Update $\lambda_k^{\mathrm{IC}}(t) \leftarrow \lambda_k^{\mathrm{IC}}(t) - (\lambda_a^{\mathrm{SM}}(t) - z_{a,k}(t))$; $\lambda_k^{\mathrm{RM}}(t) \leftarrow \lambda_k^{\mathrm{RM}}(t) + (\lambda_a^{\mathrm{SM}}(t) - z_{a,k}(t))$;
6. Else if $s_k(t) =$ Idle:

   $\lambda_k^{\mathrm{IC}}(t) \leftarrow \lambda_k^{\mathrm{IC}}(t) - \lambda_a^{\mathrm{SM}}(t)$; $\lambda_k^{\mathrm{RM}}(t) \leftarrow \lambda_k^{\mathrm{RM}}(t) + \lambda_a^{\mathrm{SM}}(t)$;

Figure 3.11: Update state information at $k$.

**State Information Update at node $a$**

1. Update $\mathcal{Y}_a(t) \leftarrow \mathcal{Y}_a(t) - \{k\}$, $\mathcal{B}_a(t) \leftarrow \mathcal{B}_a(t) \cup \{k\}$;
2. If $s_k(t) =$ Tx:
3. Update $\lambda_a^{\mathrm{IC}}(t) \leftarrow \lambda_a^{\mathrm{IC}}(t) + (\lambda_k^{\mathrm{SM}}(t) - z_{k,a}(t))$; $\lambda_a^{\mathrm{RM}}(t) \leftarrow \lambda_a^{\mathrm{RM}}(t) - (\lambda_k^{\mathrm{SM}}(t) - z_{k,a}(t))$;
4. Else if $s_k(t) =$ Rx:
5. Update $\lambda_a^{\mathrm{IC}}(t) \leftarrow \lambda_a^{\mathrm{IC}}(t) + (\lambda_k^{\mathrm{SM}}(t) - z_{a,k}(t))$; $\lambda_a^{\mathrm{RM}}(t) \leftarrow \lambda_a^{\mathrm{RM}}(t) - (\lambda_k^{\mathrm{SM}}(t) - z_{a,k}(t))$;
6. Else if $s_k(t) =$ Idle: No Changes;

Figure 3.12: Update state information at $a$.

the set of nodes in $\mathcal{Y}_k(t)$ being positioned *after* node $k$. That is, there is a relative ordering among nodes in $\mathcal{B}_k(t)$, node $k$, and nodes in $\mathcal{Y}_k(t)$. Under this notion, node $k$, being positioned after the set of nodes in $\mathcal{B}_k(t)$, is responsible to cancel interference to/from nodes in $\mathcal{B}_k(t)$. Note that we did not make a finer distinction of the relative positions (or ordering) among the set of nodes in $\mathcal{B}_k(t)$ or $\mathcal{Y}_k(t)$.

In this section, we show that the coarse set-based ordering $\mathcal{B}_k(t)$ or $\mathcal{Y}_k(t)$ at node $k$ locally can in fact be mapped into a "global node ordering" for IC among all the nodes explicitly. More formally, we give the following definition.

**Definition 3.1.** A global node ordering for IC is a list of nodes where the position of a node in the list determines its IC responsibility. Based on this list, a node is responsible for canceling interference to/from these nodes that are before itself in the list; a node does not need to cancel the interference to/from those nodes that are after itself in the list as that interference will be canceled by those nodes.

Based on this definition, we show if there exists a global node ordering for IC among the active nodes in the network, then there exists a set of feasible precoding vectors at each secondary transmitter and decoding vectors at each secondary receiver so that all data (in both primary and secondary networks) can be transported free of interference using zero-forcing technique on the secondary nodes. That is, if a global node ordering for IC exists among the active nodes, then there exist feasible precoding and decoding vectors at the PHY layer to implement the desired IC and SM in the network.

**Lemma 3.1.** *Upon the termination of the distributed algorithm, there exists a global node ordering for IC among all nodes in each time slot $t$.*

*Proof.* Before we start our algorithm, all secondary nodes are inactive and there does not exist any ordering among the nodes. Since none of the primary nodes perform IC (due to the fact that potential interference among the primary nodes is handled by interference avoidance through time slot), we can envision a list containing all active primary nodes with arbitrary order among them.

We will build upon this list to establish a global node ordering for IC.

To achieve TC, all interferences to/from primary network is canceled by the secondary nodes. Following the definition of global node ordering for IC, the secondary nodes must be placed after the primary nodes. We now show that we can maintain a global node ordering for IC at each iteration. Upon termination of the last iteration, the list remains a global node ordering for IC.

If a link fails to increase one data stream at the end of any iteration, the current global ordering for IC will not be affected. So in our proof, we only need to discuss the case that we can increase one data stream upon the end of the an iteration. Our proof is based on induction. For the first iteration, a secondary link is selected for data stream increment. We append the secondary transmit and receive nodes of this link at the end of the current list. Since there is no IC relationship between the secondary transmit and receive nodes of the chosen link, we can put transmit node either before or after the receive node. The new global ordering list consists of all active primary nodes, plus the secondary transmit and receive nodes of the chosen link. Since we can increase one data stream on this link, all interference from this link's transmit node to the neighboring primary receivers can be canceled by this transmit node, and all interference from neighboring primary transmitters to the chosen link's receive node can be canceled by this receive node. Obviously, this new list satisfies global node ordering for IC by definition.

Upon the end of $n$-th iteration, suppose there exists a global node ordering for IC. Then, we show that at the end of the $(n+1)$-th iteration, there still exists a global node ordering for IC. Denote link $(i, j)$ as the link chosen for data stream increment during the $(n+1)$ iteration. We consider two cases: (i) a data stream can be added onto $(i, j)$ without adjusting node ordering; (ii) a data stream can be added onto $(i, j)$ but requiring adjusting node ordering:

- (i) We first consider that one data stream can be added onto $(i, j)$ without adjusting node ordering in the current global node ordering list. We take node $i$ as an example. There are two cases:
  - Node $i$ is not yet on the current global node ordering list. In this case, $s_i(t)$ =Idle and

our algorithm will put node $i$'s neighboring receive nodes (not including $j$) either in $\mathcal{B}_i(t)$ or $\mathcal{Y}_i(t)$. Since all these neighboring receive nodes are active, they already have their positions in the current global ordering list in a previous iteration. If node $i$'s neighboring receive nodes are put in $\mathcal{B}_i(t)$, it is the same as putting node $i$ after these neighboring receive nodes. If node $i$'s neighboring receive nodes are put in $\mathcal{Y}_i(t)$, it is the same as putting node $i$ before these neighboring receive nodes. In either case, other nodes' relative ordering on the current global node ordering list is not affected, and thus IC responsibilities among them remain the same. Since one data stream can be added onto link $(i, j)$ successfully, node $i$ must be able to cancel interference to these nodes in $\mathcal{B}_i(t)$ (if these receive nodes are put into $\mathcal{B}_i(t)$), or the interference from $i$ can be canceled by these nodes in $\mathcal{Y}_i(t)$ (if these receive nodes are put into $\mathcal{Y}_i(t)$). Therefore, on the new list, each node is responsible for canceling interference to/from these nodes that are before itself in the list; each node does not need to cancel the interference to/from those nodes that are after itself in the list as that interference will be canceled by those nodes. By definition, the new list satisfies the global node ordering for IC.

– Node $i$ is already on the global node ordering list. In this case, $s_i(t)$=Tx and the node $i$ only performs data stream increment. There is no new node to be added to the list or none of the nodes change its position in the list. Therefore, the current ordering list is the same as that in the $n$-th iteration and satisfies the global node ordering for IC.

The case for node $j$ is similar and we omit its discussion here to conserve space.

- (ii) We then consider that one data stream can be added onto $(i, j)$ but requiring adjusting node ordering in the current global node ordering. We assume that node $k \in \mathcal{D}_{(i,j)}(t)$ is under consideration for adjustment. Suppose $k$ is a transmit node. Recall that a necessary condition for a receive node $a \in \mathcal{B}_k(t)$ to be moved to $\mathcal{Y}_k(t)$ is that: for each transmit node $x \in \mathcal{Y}_a(t)$ that needs to do IC to $a$ (i.e., $a \leftarrow x$), there cannot exist a receive node $y \in \mathcal{Y}_x(t)$ that $y$ handles IC from $x$ (i.e., $x \leftarrow y$), and $k$ handles IC to $y$ (i.e., $y \leftarrow k$) (see Figure 3.9(a)). That is, there does not exist a transmit node $x$ and a receive node $y$, such that the following

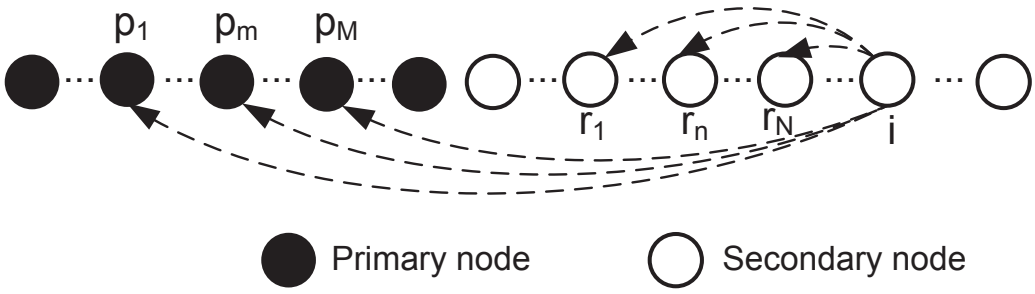



Figure 3.13: A secondary transmit node $i$ performs IC to neighboring primary and secondary receive nodes in a time slot $t$.

IC relationship holds: $a \leftarrow x \leftarrow y \leftarrow k$. Therefore, when node $a$ is chosen to move from $\mathcal{B}_k(t)$ to $\mathcal{Y}_k(t)$, node $a$'s IC responsibility for other transmit nodes, and node $k$'s IC responsibility for other receive nodes will not change, except changing the position of $k$ and $a$ (i.e., moving $a$ after $k$). Since this change is successful, $a$ is able to cancel the interference from $k$ and other transmit nodes that are before $k$. Therefore, the new list satisfies the global node ordering for IC. The discussion when $k$ is a receiver is similar.

Therefore, we conclude that after the $(n+1)$-th iteration, the new list satisfies the global node ordering for IC.

Based on the above discussion, we conclude that upon the termination of the distributed algorithm, we have a global node ordering for IC. □

**Theorem 1.** *There exists a set of feasible precoding vectors at each secondary transmitter and a feasible set of decoding vectors at each secondary receiver so that all data (in both primary and secondary networks) can be transported free of interference based on the global node ordering for IC.*

*Proof.* We first consider a secondary transmit node $i$ on the global node ordering list, as shown in Figure 3.13. The dashed arrows represent the interference from node $i$ to those receive nodes that are before node $i$ on the global node ordering list. The nodes $p_1 \cdots p_M$ are $i$'s neighboring primary

receivers, while nodes $r_1 \cdots r_N$ are $i$'s neighboring secondary receivers. Suppose that $i$ transmits $z_{(i,j)}$ data streams to secondary node $j$. Denote $\mathbf{u}_i^q$ as an $A_i \times 1$ transmit weight vector at $i$ for each data stream $q$ ($1 \leq q \leq z_{(i,j)}$), and $\mathbf{v}_j^q$ as an $A_j \times 1$ receive weight vector at receiver node $j$ to receive data stream $q$. Since each primary link only transmits one data stream, we use $\mathbf{u}_p^1$ and $\mathbf{v}_p^1$ to denote the primary node $p$'s ($p \in \{p_1, \cdots, p_M\}$) transmit and receive vectors .

Denote $\mathbf{H}_{(i,b)}$ ($b \in \{p_1, \cdots, p_M, r_1, \cdots, r_N\}$) as the $A_i \times A_b$ channel gain matrix between nodes $i$ and $b$. We assume a rich scattering environment, where all channels $H_{i,b}$ have full rank and independent with each other. To successfully transmit $z_{(i,j)}$ data stream from node $i$ to its intended receive node $j$, the transmit node $i$ should cancel all its interference to primary receive nodes $p_1$ to $p_M$ and secondary receive nodes $r_1$ to $r_N$. Then, we should have the following constraints:

$$(\mathbf{u}_i^q)^T \mathbf{H}_{(i,j)} \mathbf{v}_j^q = 1 \,, \qquad (1 \leq q \leq z_{(i,j)}) \,, \tag{3.4.1}$$

$$(\mathbf{u}_i^q)^T \mathbf{H}_{(i,j)} \mathbf{v}_j^d = 0 \,, \qquad (1 \leq q, d \leq z_{(i,j)}, d \neq q) \,, \tag{3.4.2}$$

$$(\mathbf{u}_i^q)^T \mathbf{H}_{(i,p_m)} \mathbf{v}_{p_m}^1 = 0 \,, \qquad (1 \leq q \leq z_{(i,j)}, 1 \leq m \leq M) \,, \tag{3.4.3}$$

$$(\mathbf{u}_i^q)^T \mathbf{H}_{(i,r_n)} \mathbf{v}_{r_n}^d = 0 \,, \qquad (1 \leq q \leq z_{(i,j)}, 1 \leq n \leq N, 1 \leq d \leq z_{(s_n,r_n)}) \,, \tag{3.4.4}$$

where $s_n$ is the transmit node which transports $z_{(s_n,r_n)}$ data streams to secondary receive node $r_n$.

The number of constraints from (4.5.1) and (4.5.2) are $(z_{(i,j)})^2$. The number of constraints from (4.5.3) is $z_{(i,j)} \sum_{m=1}^{M} 1$. The number of constraints from (4.5.4) is $z_{(i,j)} \sum_{n=1}^{N} z_{(s_n,r_n)}$. The total number of constraints is therefore $(z_{(i,j)} \sum_{m=1}^{M} 1 + (z_{(i,j)})^2 + z_{(i,j)} \sum_{n=1}^{N} z_{(s_n,r_n)})$. Recall that in our algorithm, in either step 2 or step 3, the total number of DoF consumption cannot be more than the total number of DoFs at a node. We have, $z_{(i,j)} \sum_{m=1}^{M} 1 + (z_{(i,j)})^2 + z_{(i,j)} \sum_{n=1}^{N} z_{(s_n,r_n)} \leq z_{(i,j)}(\sum_{m=1}^{M} 1 + z_{(i,j)} + \sum_{n=1}^{N} z_{(s_n,r_n)}) \leq z_{(i,j)} A_i$. That is, the total number of constraints is no more than $z_{(i,j)} A_i$.

Since the precoding vector $\mathbf{u}_i^q$ is an $A_i \times 1$ vector for each data stream $q$ ($1 \leq q \leq z_{(i,j)}$), the total number of variables at the transmit node $i$ is $z_{(i,j)} A_i$ and the number of variables is no less than the number of constraints. On the other hand, since the channels $H_{(i,b)}$ are full rank and independent with each other, it can be shown that the constraints in (4.5.1), (4.5.2), (4.5.3), and

(4.5.4) are linearly independent with each other based on [59]. So for any given $\mathbf{v}_j^q$ $(1 \leq q \leq z_{(i,j)})$, we are guaranteed to construct feasible precoding vectors $\mathbf{u}_i^q$ $(1 \leq q \leq z_{(i,j)})$ at transmit node $i$.

The proof that we can construct the feasible decoding vectors $\mathbf{v}_j^q$ $(1 \leq q \leq z_{(i,j)})$ at the secondary receive node $j$ (for any given precoding vectors $\mathbf{u}_i^q$ $(1 \leq q \leq z_{(i,j)})$) is similar to the transmit node $i$. We omit its discussion here to conserve space. Based on the above discussions, there exist feasible precoding/decoding vectors at the secondary nodes. Therefore, there exists a set of feasible precoding vectors at each secondary transmitter and a feasible set of decoding vectors at each secondary receiver so that all data (in both primary and secondary networks) can be transported free of interference. This completes the proof. □

## 3.5 Complexity Analysis

We now show that the overall computation complexity of the distributed algorithm is polynomial time. Step 1 (ranking of active secondary links) is done only once. As shown in [81], this step can be done in $O(S^2)$. The iteration of our algorithm involves steps 2 and 3. We now analyze the complexity of each iteration and the number of iterations required in the algorithm.

In step 2, nodes $i$ and $j$ (for a chosen link $(i, j)$) need to check the feasibility of increasing one more data stream over at most $T$ time slots. The worst case scenario is that both nodes $i$ and $j$ are idle (case (a) in Fig. 3.4). Since both nodes $i$ and $j$ need to check the number of remaining DoFs of each of their neighboring nodes and the number of DoFs used for SM by these nodes, the complexity of this operation is $O(2S)$. Afterward, nodes $i$ and $j$, and their neighbors, need to update their DoF allocation status. The complexity of this operation is $O(S)$. Since there is a total of $T$ time slots, the total complexity of this step is $T \cdot O(2S + S) = O(ST)$.

In step 3, nodes $i$ and $j$, as well as their neighboring nodes attempt to adjust IC responsibility in at most $T$ time slots. During each time slot, the computation consists of three parts: (i) identifying the subset of nodes $\mathcal{D}_{(i,j)}(t)$, which has a complexity $O(S)$; (ii) identifying the set of nodes $\bar{\mathcal{B}}_k(t)$ for each $k \in \mathcal{D}_{(i,j)}(t)$, which has a complexity of $O(S^2)$. Since the number of nodes in $\mathcal{D}_{(i,j)}(t)$ is

at most $S$, the total complexity for (ii) is $O(S*S^2) = O(S^3)$; (iii) adjusting the IC responsibility for each node $k \in \mathcal{D}_{(i,j)}(t)$, and updating each node's state information, which has a complexity $O(S)$. Since there is a total of $T$ time slots, the total complexity of step 3 is $T \cdot O(S + S^3 + S) = O(TS^3)$.

Since each node has $A$ antennas and there are $L$ active links in the network, the number of iterations of our algorithm is at most $O(LA)$. Therefore the overall complexity is $O(S^2 + O(LA) \cdot [O(ST) + O(TS^3)]) = O(LATS^3)$.

## 3.6 Simulation Results

In this section, we present simulation results to demonstrate the performance of the proposed distributed algorithm. We compare our results with the centralized methods as discussed in [77]. Since the centralized problem formulation is MILP, which is NP-hard in general, we cannot obtain the optimal solution for comparison. Instead, we will compare the performance of our algorithm to an upper bound of the objective for the centralized problem. Such an upper bound can be obtained by running CPLEX for a given termination time. Clearly, such a comparison approach is very aggressive and conservation. This is because the optimal objective value (not obtainable) to the centralized problem lies between the upper bound and the feasible solution obtained by our distributed algorithm. Therefore, if the feasible solution from our distributed algorithm is somehow close to the upper bound by CPLEX, then we can claim that our solution (objective) is even closer to the optimal objective and thus is competitive.

### 3.6.1 Simulation Setting

We consider a secondary CR network co-locates with a primary network within a $100 \times 100$ area. For generality, we normalize the units for distance, bandwidth, and data rate with appropriate dimensions. Each node (both primary and secondary) is randomly deployed inside the $100 \times 100$ area. The primary nodes are traditional single-antenna node while the secondary nodes are

equipped with MIMO, with four antennas on each node. We assume that each node's transmission range and interference range are 30 and 50, respectively. We assume a time frame is divided into $T = 10$ time slots.

### 3.6.2 A Case Study

Before we present complete results, we show results for one network instance, with 20 primary nodes and 30 secondary nodes. The location of each node is listed in Figure 3.14. We assume there are three primary sessions and four secondary sessions, with each session's source and destination nodes shown in Figure 3.14. For simplicity, we assume that minimum-hop routing is used for each primary and secondary sessions, although other routing methods may be used if needed. Figure 3.14 shows the routing topology for each primary and secondary sessions. where a solid line represents a primary link and a dashed line represents a secondary link. Scheduling for the primary and secondary links is given in this figure, where numbers in the box represent the time slots used by the corresponding link. Note that scheduling for the primary links is solely determined by the primary network, while scheduling for each secondary link is found by our distributed algorithm.

The objective value obtained from our distributed algorithm is 0.6 (in less than a second computational time). On the other hand, the upper bound obtained by CPLEX is 0.7 (with a cut-off time of 8 hours). As discussed, since the optimal solution lies between 0.6 and 0.7, our objective value (0.6) is quite close to the unknown optimal.

To show whether TC is achieved by the secondary network, we consider one time slot, say 6. Figure 3.15 shows the set of active links in time slot 6 for both networks. In this time slot, secondary links $S_{28} \to S_{17}, S_{13} \to S_{24}, S_{30} \to S_{12}, S_3 \to S_1, S_4 \to S_{11}$ and $S_4 \to S_5$ are active simultaneous with primary links $P_1 \to P_8$ and $P_4 \to P_9$, through IC by the secondary nodes.

We first consider inter-network IC:

- For secondary link $S_{28} \to S_{17}$, its interference to $P_9$ on primary link $P_4 \to P_9$ is canceled by $S_{28}$ with 1 DoF, while the interference from $P_4$ and $P_1$ to $S_{17}$ is canceled by $S_{17}$, each with

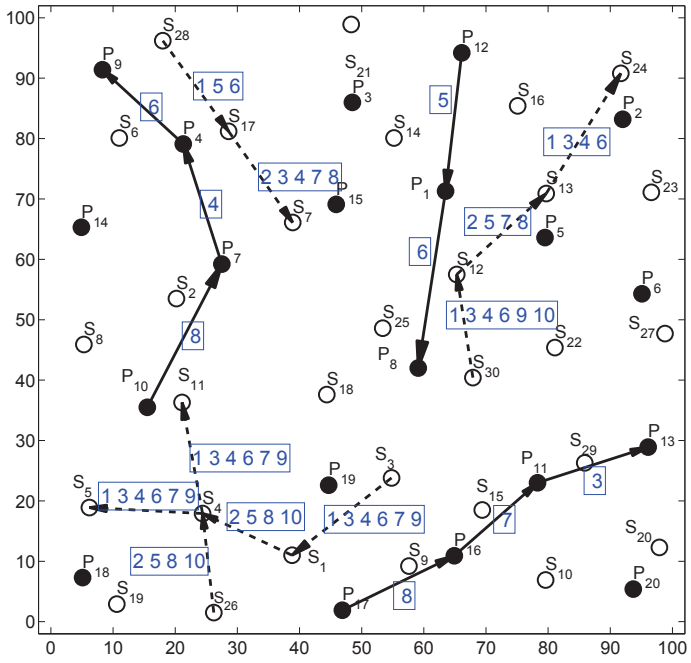



Figure 3.14: Routing topology for each primary and secondary sessions and scheduling on each link of the respective route. The numbers in the box next to a link show the time slots when the link is active.

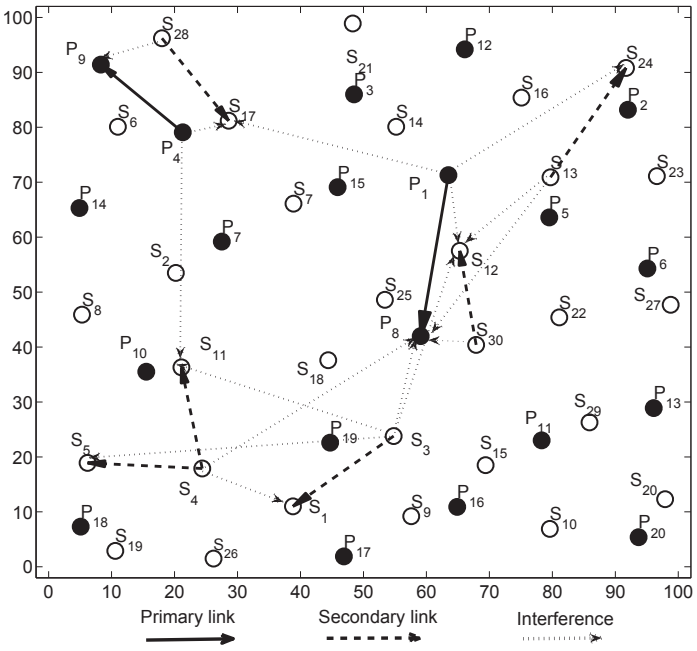



Figure 3.15: Active links in time slot 6 in both primary and secondary networks.

1 DoF.

- For secondary links $S_3 \rightarrow S_1, S_{30} \rightarrow S_{12}, S_{13} \rightarrow S_{24}, S_4 \rightarrow S_{11}$ and $S_4 \rightarrow S_5$, the interference from their transmitters $(S_3, S_{30}, S_{13}, S_4)$ to receiver $P_8$ on primary link $P_1 \rightarrow P_8$ is canceled by $S_3, S_{30}, S_{13}$ and $S_4$, each with 1 DoF. The interference from $P_1$ to $S_{12}$ and $S_{24}$ is canceled by $S_{12}$ and $S_{24}$ with 1 DoF, respectively, and the interference from $P_4$ to $S_{11}$ is canceled by $S_{11}$ with 1 DoF.

For intra-network IC within the secondary network, our solution shows that:

- $S_{11}$ is canceling interference from $S_3$ and $S_4$, each with 1 DoF.
- $S_5$ is canceling interference from $S_3$ and $S_4$, each with 1 DoF.
- The interference from $S_4$ to $S_1$ is canceled by $S_1$ with 1 DoF.
- The interference from $S_3$ to $S_{12}$ is canceled by $S_{12}$ with 1 DoF.
- The interference from $S_{13}$ to $S_{12}$ is canceled by $S_{13}$ with 2 DoFs.
- The interference from $S_{30}$ to $S_1$ and $S_{11}$ is canceled by $S_{30}$, each with 1 DoF.

The details of DoF allocation for SM and IC at each active secondary node in time slot $6$ are shown in Table 3.2. In this table, the second and third columns represent the set of secondary nodes that are in $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$ (i.e., before and after this node in the global node ordering) in our distributed algorithm, respectively. The fourth column represents the number of DoFs allocated for SM. The fifth column represents the number of DoFs that are allocated for IC to/from primary network. The last column represents the number of DoFs allocated for IC for the set of secondary nodes in $\mathcal{B}_i(t)$.

Now, we show that there exists a global node ordering for IC among all nodes in time slot $6$. Based on Table 3.2, we can establish a global node ordering for IC among all nodes explicitly. Since none of the primary nodes perform IC, we put active primary nodes $p_1$, $p_4$, $p_8$ and $p_9$ in the

Table 3.2: DoF allocation for SM and IC at each active secondary node in time slot 6.

| Node $i$ | $\mathcal{B}_i(t)$ | $\mathcal{Y}_i(t)$ | DoF for SM | IC to/from primary | DoF for IC within secondary network |
|---|---|---|---|---|---|
| $S_1$ | $\{S_4\}$ | $\{S_{30}\}$ | 1 | 0 | 2 |
| $S_3$ | | $\{S_5, S_{11}, S_{12}\}$ | 1 | 1 | 0 |
| $S_4$ | | $\{S_1, S_5, S_{11}\}$ | 2 | 1 | 0 |
| $S_5$ | $\{S_3, S_4\}$ | | 1 | 0 | 2 |
| $S_{11}$ | $\{S_3, S_4\}$ | $\{S_{30}\}$ | 1 | 1 | 2 |
| $S_{12}$ | $\{S_3\}$ | $\{S_{13}\}$ | 1 | 1 | 1 |
| $S_{13}$ | $\{S_{12}\}$ | | 1 | 1 | 2 |
| $S_{17}$ | | | 1 | 2 | 0 |
| $S_{24}$ | | | 1 | 1 | 0 |
| $S_{28}$ | | | 1 | 1 | 0 |
| $S_{30}$ | $\{S_1, S_{11}\}$ | | 1 | 1 | 2 |

front of global node ordering list with arbitrary order among them. Based on $\mathcal{B}_i(t)$ and $\mathcal{Y}_i(t)$ in Table 3.2, we can establish a global ordering among the secondary nodes, as shown in Figure 3.16. The arrows originating from a node in the figure represent the interference from that node.

In this figure, we first take a receive node $S_{12}$ as an example. $S_{12}$ is being interfered by transmit nodes $P_1$, $S_3$ and $S_{13}$. Since $P_1$ and $S_3$ are before $S_{12}$, $S_{12}$ is responsible for canceling their interference, each with 1 DoF. For the interference from $S_{13}$, $S_{12}$ does not need to use any DoF to cancel this interference, since $S_{13}$ is after $S_{12}$ in this global node ordering. This interference is to be canceled by $S_{13}$ with 2 DoFs. As a second example, consider transmit node $S_3$. $S_3$ is interfering receive nodes $P_8$, $S_5$, $S_{11}$ and $S_{12}$. Since $P_8$ is before $S_3$, $S_3$ is responsible for canceling this interference with 1 DoF. For its interference to $S_5$, $S_{11}$ and $S_{12}$, $S_3$ does not need to use any DoF to cancel this interference, since $S_5$, $S_{11}$ and $S_{12}$ are after $S_3$ in this global node ordering. This interference is canceled by $S_5$, $S_{11}$ and $S_{12}$, respectively, each with 1 DoF. It is easy to verify that

based on this global node ordering, the IC responsibilities at nodes $S_4$, $S_{17}$, $S_{24}$, $S_{28}$, $S_5$, $S_{11}$, $S_1$, $S_{30}$ and $S_{13}$ are all satisfied.

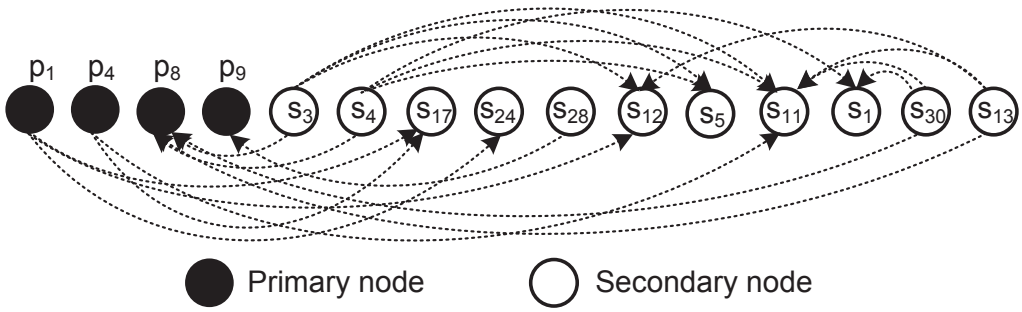



Figure 3.16: A global node ordering for IC in time slot 6.

### 3.6.3 Comparison to Interweave Paradigm

To show the benefits of TC paradigm, we compare our results to those under the interweave paradigm. For the latter, a secondary node is not allowed to transmit (receive) at the same time when a nearby primary node is active. That is, the secondary nodes will not perform inter-network IC for interference to/from the primary nodes. The problem formulation for this paradigm is given in [78], which is similar to the problem formulation for TC paradigm except that we remove DoF allocation by the secondary nodes to cancel interference to/from the primary nodes. The problem formulation remains an MILP, and an upper bound can be obtained by running CPLEX for a given termination time (i.e., 8 hours).

Following the same setting as in the case study in the last section, we obtain an upper bound of 0.4 for the objective value. (comparing to 0.6 from our distributed solution in Section 3.6.2). The time slot scheduling on each link of the secondary sessions is shown in Fig. 3.17. Comparing Fig. 3.14 and 3.17, we find that the set of time slots used by each secondary link under interweave paradigm is smaller. We take the link $S_{28} \rightarrow S_{17}$ as an example. Under interweave paradigm, this link cannot use time slot 6 as the neighboring primary link $P_4 \rightarrow P_9$ is using this it. However under TC paradigm, this link can use time slot 6 to achieve the simultaneous activation with the primary link $P_4 \rightarrow P_9$. For any secondary link in Figure 3.17, we cannot find one that simultaneously

actives with the primary links. There is no inter-network interference in the network.

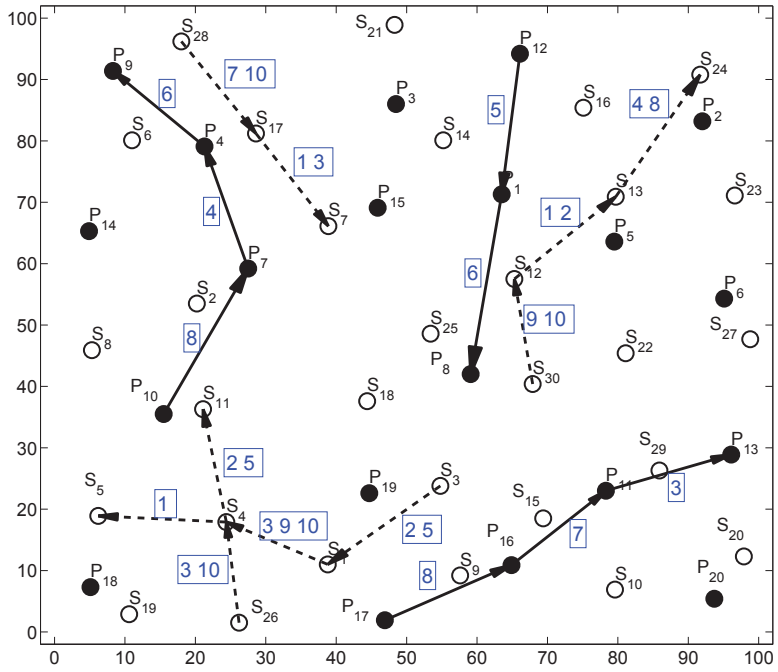



Figure 3.17: Routing for each session and scheduling on each link for both primary and secondary networks under the interweave paradigm.

### 3.6.4 Complete Results

We run our distributed algorithm for 50 random network instances, with 20-node primary network and 30-node secondary network. The number of primary and secondary sessions are random, with the source and destination nodes of each session are randomly generated. Table 3.3 compares the objective values from our distributed algorithm and the upper bounds from CPLEX solver. The average ratio between the two over 50 instances is 83.7%, with standard derivation of 0.073. Since the optimal objective value (unknown) to the centralized problem lies between the upper bound and the feasible solution obtained by our distributed algorithm, these results affirm that our distributed algorithm is highly competitive.

Table 3.3: Results for 50 network instances.

| Instance | Our Algorithm | CPLEX | Instance | Our Algorithm | CPLEX |
|---|---|---|---|---|---|
| 1 | 0.8 | 0.9 | 26 | 0.7 | 0.8 |
| 2 | 0.7 | 0.9 | 27 | 0.6 | 0.7 |
| 3 | 0.4 | 0.5 | 28 | 0.5 | 0.7 |
| 4 | 0.4 | 0.4 | 29 | 0.5 | 0.6 |
| 5 | 0.4 | 0.6 | 30 | 0.7 | 0.8 |
| 6 | 0.5 | 0.6 | 31 | 1.0 | 1.1 |
| 7 | 0.9 | 1.1 | 32 | 0.8 | 1.0 |
| 8 | 0.7 | 0.8 | 33 | 0.3 | 0.4 |
| 9 | 1.1 | 1.1 | 34 | 0.7 | 0.9 |
| 10 | 0.3 | 0.3 | 35 | 0.5 | 0.6 |
| 11 | 0.6 | 0.7 | 36 | 0.8 | 0.9 |
| 12 | 0.7 | 0.8 | 37 | 0.6 | 0.8 |
| 13 | 0.3 | 0.4 | 38 | 0.5 | 0.5 |
| 14 | 0.9 | 1.0 | 39 | 0.6 | 0.7 |
| 15 | 0.7 | 0.8 | 40 | 0.4 | 0.4 |
| 16 | 1.0 | 1.0 | 41 | 0.6 | 0.7 |
| 17 | 0.9 | 1.0 | 42 | 0.8 | 0.9 |
| 18 | 0.2 | 0.4 | 43 | 0.3 | 0.3 |
| 19 | 0.6 | 0.6 | 44 | 0.5 | 0.7 |
| 20 | 0.6 | 0.7 | 45 | 0.4 | 0.5 |
| 21 | 1.1 | 1.1 | 46 | 0.6 | 0.6 |
| 22 | 0.6 | 0.7 | 47 | 0.8 | 0.9 |
| 23 | 0.8 | 0.8 | 48 | 0.4 | 0.5 |
| 24 | 0.6 | 0.9 | 49 | 0.5 | 0.6 |
| 25 | 0.6 | 0.6 | 50 | 0.8 | 1.0 |

## 3.7 Chapter Summary

TC is a new spectrum sharing paradigm that allows simultaneous activation of the secondary nodes with the primary nodes. The enabling PHY layer technology for TC is IC, which is the sole responsibility of the secondary nodes. In this chapter, we design a distributed algorithm to achieve TC for multi-hop primary and secondary networks. The main challenge in this algorithm is to ensure that IC is done efficiently (i.e., canceled once by a secondary node) and in a feasible manner (i.e., implementable at the PHY layer). In contrary to a centralized IC algorithm which relies on a global node ordering, we only maintain two local sets for each node to keep track of the node's IC responsibilities. We show how to establish, maintain, and update these two local sets for each node in each iteration of our distributed algorithm. Our distributed algorithm increases the data stream on each active link iteratively based on local computation. Since the nodes in the two local set of a node directly affect the node's IC responsibility, our algorithm attempts to switch nodes in the two sets if it can improve the IC structure. Although no explicit node ordering is maintained in our distributed algorithm, we prove that our distributed data structure at each node (with the use of two local sets) can be mapped to an explicit global node ordering for IC among all nodes in the network. This guarantees the existences of feasible precoding/decoding vectors at the secondary nodes to achieve our desired IC in the network (i.e., feasibility at the PHY layer). Through simulation study, we show that our distributed algorithm achieves TC between secondary and primary networks and offers competitive throughput performance when compared to a centralized optimization.

# Chapter 4

# Transparent Coexistence: An Online Algorithm

## 4.1 Introduction

There has been extensive research on exploring coexistence between primary and secondary networks in recent years. In [22], Goldsmith *et al.* identified three coexistence paradigms, namely *interweave, underlay,* and *overlay.* The interweave paradigm follows the traditional interference avoidance, which refers to that the secondary nodes are allowed to use a spectrum allocated to the primary nodes only when the primary nodes do not use it (in time, frequency, or space) [21,26,72]. In this way, interference is effectively avoided through *interweaving* spectrum access between primary and secondary nodes. On the other hand, the underlay paradigm refers to that the secondary nodes are allowed to be active simultaneously on the same spectrum with the primary nodes, as long as the interference produced by the secondary nodes are controlled properly (e.g., through effective interference cancelation [23, 33, 85, 86]). The overlay paradigm refers to that there are some levels of cooperation between the primary and secondary nodes in data forwarding [31,42,61,79,83].

One of the biggest challenges for all three paradigms is how to handle the dynamic changes for online traffic arrival and departure in both the primary and secondary networks. Typically, a secondary session arrives and departs over time and so does a primary session. The problem is particularly difficult in a distributed multi-hop network environment. This is because, when a new primary or secondary session arrives, one must quickly make an online decision on whether or not the new session can be admitted into the network. This problem is addressed differently under each of the three paradigms, each with its own unique challenges and solutions. In this chapter, we attempt to address this problem for the underlay paradigm, which we believe is the most difficult among the three. This is because unlike overlay, underlay does not allow active cooperation between the primary and secondary nodes and puts all burden related to interference management to the secondary nodes. Also, unlike interweave, underlay allows simultaneous activation of the secondary nodes with the primary nodes through interference cancellation (IC), which is much more aggressive and complex than merely avoiding interference.

There were active efforts to study efficient online algorithms to handle traffic dynamics even in the old days for the telephone network. But the problems there were much simpler (e.g., wired network, no consideration of IC). For spectrum sharing in the interweave paradigm, there have been some recent studies on the handling of dynamic traffic (see, e.g., [16, 25, 35]). The focus there was mainly on efficiently utilizing spectrum holes and to avoid interference to the primary users (no active IC). In the overlay paradigm, there have also been some studies addressing dynamic traffics (see, e.g., [17, 70]. The primary goal here is to identify optimal scheduling so that traffic can be successfully relayed cooperatively from a source to its destination node. In the underlay paradigm, the problem becomes much harder as the goal is to enable aggressive (simultaneous) spectrum access by the secondary nodes through IC to the primary nodes. To our knowledge, there has not been much work on how to handle traffic dynamics in the underlay paradigm.

The goal of this chapter is to design a fast online algorithm to handle dynamics session arrival and departure in the underlay paradigm. As discussed, algorithms to handle traffic dynamics in the underlay paradigm is likely the most challenging among the three paradigms due to IC. For IC, we consider to employ multiple antennas on the secondary nodes. Since it takes time to configure the

precoding/decoding vectors at a secondary node for spatial multiplexing (SM) and IC, per packet level dynamic traffic management does not appear to be practical. Instead, our traffic management algorithm is to address session (flow) level dynamics, i.e., to determine if a new session can be admitted into the network and how to control the additional IC that comes with it. The main contribution of this chapter is an online distributed algorithm to handle session-level dynamics for the underlay paradigm. In particular, our algorithm is designed with the following capabilities and features:

- When a new secondary session initiates, the algorithm is able to make a quick decision on whether or not it can join the network through distributed computation. If a secondary session is admitted into the network, then our algorithm will configure MIMO degree-of-freedom (DoFs) at each secondary node so that all interference to/from the primary nodes are properly canceled, as required for underlay coexistence.

- When a new primary session enters the network, the algorithm is able to vacate any active secondary session that may be of hinderance. An active secondary session is allowed to be active only if they are able to cancel all interference to/from the primary nodes.

- At all time, our algorithm is able to guarantee that IC (as defined by MIMO DoF allocation) is feasible at the PHY layer for all MIMO transmitters and receivers. By "feasible" at the PHY layer, we mean that there exist a set of feasible precoding vectors at the secondary transmitters and a set of feasible decoding vectors at the secondary receivers at the PHY layer so that all data (in both primary and secondary networks) can be transported free of interference.

- Our online distributed algorithm is able to offer competitive performance when compared to an offline centralized algorithm.

The remainder of this chapter is organized as follows. In Section 4.2, we review the underlay coexistence paradigm and understand how interference is managed at the PHY layer. In Section 4.3, we describe network setting and discuss the problem that we are going to study in this

chapter. In Section 4.4, we propose an online distributed algorithm to handle initiation and termination of primary/secondary sessions in the underlay coexistence paradigm. A proof of PHY layer feasibility of our algorithm is also given in Section 4.5. Section 7.6 presents performance evaluation of our algorithm. Section 7.9 concludes this chapter.

## 4.2 Transparent Coexistence: A Primer

The underlay coexistence paradigm refers to that the secondary network is allowed to be active concurrently with the primary network on the same spectrum, as long as its interference to the primary network is negligible (e.g., kept at the noise floor) [22]. In contrast to interweave, which solely relies on interference avoidance, underlay relies on more powerful interference management techniques to enable concurrent activations of both primary and secondary networks. In underlay, the primary nodes' behavior is *not* affected by the secondary nodes. The primary nodes may use the spectrum freely to serve their needs as if they were the only nodes that use the spectrum. On the other hand, to ensure their interference to the primary nodes is negligible, the secondary nodes need to take appropriate measures in interference management during their transmissions. To ensure "underlay", all burdens (or activities) on interference management must rest solely on the secondary nodes and remain unnoticeable to the primary nodes.

There are many measures that the secondary nodes can take to control its interference to the primary network. These measures are typically done at the physical layer. The most simple approach is to have secondary transmitters to meticulously control its output power so that the received power at neighboring primary receivers remains below some prescribed interference threshold (e.g., noise floor). Since such interference threshold is typically low, the secondary nodes' transmit power, therefore, must be kept at a very low level, resulting in a seriously limitation in coverage and connectivity. Another approach for the secondary nodes to control its interference is to use UWB, with which the secondary nodes can spread their signals over a wide bandwidth (so that they are below the noise floor) and then despread the wideband signals at secondary receivers. This technique has

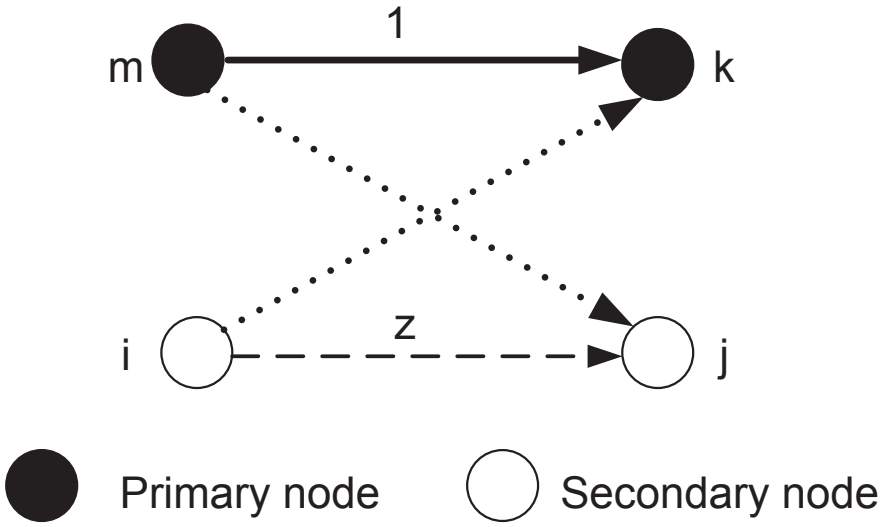



Figure 4.1: The underlay coexistence of one secondary link with one primary link. A solid line represents a primary link, a dashed line represents a secondary link, and a dotted line represents a interference.

been used in [82]. The limitation of UWB is that it requires wide bandwidth and allows only short range transmission. The most promising approach for the secondary nodes to control its interference, as we perceive, is to exploit IC capabilities offered by multiple antennas at the node (i.e., MIMO). MIMO has already become pervasive in wireless communications (e.g., cellular, WiFi) and offers unprecedented capabilities in improving throughput, mitigate interference, and enhancing reliability [8, 68]. There have been some active efforts on exploiting MIMO on the secondary nodes in the underlay paradigm [23, 33, 75, 85, 86].

To understand how MIMO can help the secondary nodes achieve underlay coexistence paradigm, we consider the following simple example. In Fig. 4.1, we have a pair of primary transmit/receive nodes and a pair of secondary transmit/receive nodes. Suppose the primary transmit/receive nodes ($m$ and $k$) are each equipped with a single antenna, while the secondary transmit/receive nodes ($i$ and $j$) are each equipped with four antennas. We assume the primary node $m$ is transmitting one data stream to the primary receive node $k$. To allow concurrently transmission from secondary node $i$ to node $j$, we must ensure that the interference from node $i$ is canceled at primary receiver

$k$ so that $k$ does not feel the presence of the secondary ndoes' activities. Further, at secondary receive node $j$, the interference from primary transmitter $m$ must be canceled. Otherwise, node $j$ will not be able to decode the signals from node $i$.

In this example (Fig. 4.1), we assume that secondary node $i$ hopes to transmit $z$ data streams to secondary receive node $j$. For data stream $a = 1, \cdots, z$, denote $\mathbf{u}_i^a$ as its $4 \times 1$ transmit vector at node $i$ and $\mathbf{v}_j^a$ as a $4 \times 1$ receive vector at receive node $j$. For the data stream from primary transmit node $m$ to receive node $k$, denote $u_m$ and $v_k$ as the weights at transmit node $m$ and receive node $k$, respectively. Denote $\mathbf{H}_{(i,j)}$, $\mathbf{H}_{(i,k)}$, and $\mathbf{H}_{(m,j)}$ as the channel matrices between node $i$ and $j$, $i$ and $k$, and $m$ and $j$, respectively. The dimensions of $\mathbf{H}_{(i,j)}$, $\mathbf{H}_{(i,k)}$, and $\mathbf{H}_{(m,j)}$ are $4 \times 4$, $4 \times 1$, and $1 \times 4$, respectively. We assume all channels are of full rank. To achieve underlay, secondary transmit node $i$ must cancel its interference to primary receive nodes $k$. We have

$$(\mathbf{u}_i^a)^T \mathbf{H}_{(i,k)} v_k = 0 \ , \qquad (1 \leq a \leq z) \ . \tag{4.2.1}$$

In addition, to have secondary node $j$ to receive from $i$ free of the interference from primary transmit node $m$, secondary node $j$ must cancel this interference. We have

$$u_m \mathbf{H}_{(m,j)} \mathbf{v}_j^a = 0 \ , \qquad (1 \leq a \leq z) \ . \tag{4.2.2}$$

After canceling all interference to the primary receiver and from the primary transmitter, the secondary transmit node $i$ may transmit $z$ data streams to its intended receive node $j$ via spatial multiplexing (SM). We have:

$$(\mathbf{u}_i^a)^T \mathbf{H}_{(i,j)} \mathbf{v}_j^a = 1 \ , \qquad (1 \leq a \leq z) \ , \tag{4.2.3}$$

$$(\mathbf{u}_i^a)^T \mathbf{H}_{(i,j)} \mathbf{v}_j^b = 0 \ , \qquad (1 \leq a \leq z, 1 \leq b \leq z, a \neq b) \ . \tag{4.2.4}$$

If we can find a feasible solution to $\mathbf{u}_i^a$ and $\mathbf{v}_j^a$ for (4.2.1), (4.2.2), (4.2.3), and (4.2.4), then the secondary link ($i$ to $j$) can be active at the same time as the primary link and thus we can achieve underlay for the secondary link.

We now show that we can indeed find a feasible solution to $\mathbf{u}_i^a$ and $\mathbf{v}_j^a$ for (4.2.1), (4.2.2), (4.2.3), and (4.2.4). Let's consider the first data stream. In constraint (4.2.2), since $u_m$ is a constant,

$\mathbf{H}_{(m,j)}$ is a $1 \times 4$ constant matrix, one can always find $z$ ($z \leq 4$) feasible vectors $\mathbf{v}_j^1, \cdots, \mathbf{v}_j^z$ that satisfy this constraint. Now suppose we use one such set of feasible vectors, i.e., $\mathbf{v}_j^1, \cdots, \mathbf{v}_j^z$ are now constant vectors. For precoding vector $\mathbf{u}_i^1$ (a $4\times1$ vector with 4 free variables), it is constrained by (4.2.1), (4.2.3), and (4.2.4), which has a total of $(1+z)$ constraints. If $1+z \leq 4$, the number of constraints is no more than the number of variables, then there always exists a feasible precoding vector $\mathbf{u}_i^1$ satisfying (4.2.1), (4.2.3), and (4.2.4). That is, as long as $z \leq 3$, we can find a feasible $\mathbf{u}_i^1$. The same arguments hold for $\mathbf{u}_i^2, \cdots, \mathbf{u}_i^z$. That is, for $z \leq 3$, we can construct a set of feasible precoding vectors $\mathbf{u}_i^1, \cdots, \mathbf{u}_i^z$ and decoding vectors $\mathbf{v}_j^1, \cdots, \mathbf{v}_j^z$ that achieve the desired IC (at nodes $k$ and $j$) and SM (from $i$ to $j$).

Instead of working with complex matrix representation, a simple model to quantify MIMO resources at a node is the so-called degree-of-freedom (DoF) [30,68]. Simply put, the total number of DoFs at a node (no more than the number of antenna elements) represents the available resource at the node. A DoF can be used for either data transmission/receiption or IC. Typically, for SM, transmitting one data stream requires one DoF at the transmitter and one DoF at the receiver. For IC to/from the primary network, the number of DoFs required at a secondary transmitter is equal to the number of data streams that are received at the neighboring primary receivers, while the number of DoFs required at a secondary receiver for IC is equal to the number of data streams that are transmitting at the neighboring primary transmitters. The total number of the DoFs consumption (for SM and IC) cannot be more than the number of antennas. For the simple example in Fig. 4.1, the primary transmitter $m$ uses 1 DoF to transmit 1 data stream to its receiver $k$. The secondary nodes $i$ and $j$ each has 4 DoFs. Secondary transmitter $i$ uses 1 DoF to cancel its interference to primary receiver $k$ (as $k$ is receiving 1 data stream from $m$). Secondary receiver $j$ uses 1 DoF to cancel the interference from primary transmitter $m$ (as $m$ is transmitting 1 data stream). Now node $i$ and $j$ each has 3 DoFs left and can transmit up to 3 data streams from $i$ to $j$.

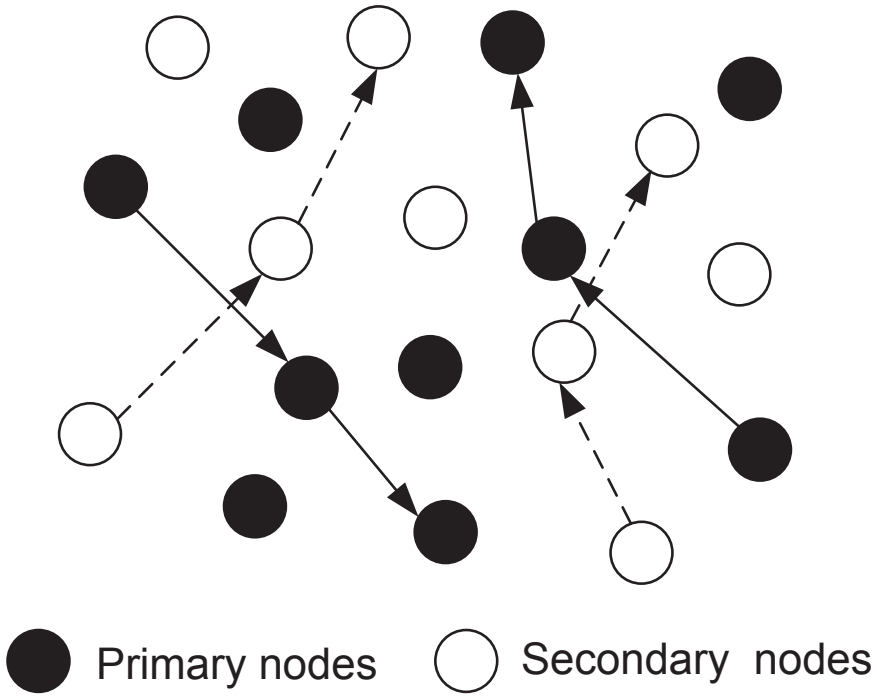



Figure 4.2: The multi-hop primary and secondary networks.

## 4.3 Problem Statement

In this chapter, we address underlay coexistence for the secondary users under dynamic traffic patterns. Consider a set of primary nodes $\mathcal{P}$ co-locate with a set of secondary nodes $\mathcal{S}$ in the same geographical region. Within the primary network, new sessions arrive following a Poisson process. Each new session consists of a source and a destination node and employs shortest path (unicast) routing (e.g., AODV [45], DSR [32]), as shown in Fig. 4.2. If the new primary session can be supported (through time slot scheduling), its holding time will follow certain distribution. Upon completing its holding time, the primary session will terminate and leave the network. Given that the primary nodes do not have any IC responsibility, we assume each primary node is equipped with a single antenna (just as in Fig. 4.1). For a new primary session, we assume it has a rate requirement of 1 data stream, which can be supported on a single antenna. For scheduling, suppose there are $T$ time slots in a frame. The primary nodes can use this set of time slots freely as if they are the only nodes in the network (without any consideration of the secondary nodes). To ensure

mutual and self interference are avoided, we need to have a feasible scheduling solution for all active primary sessions. If a new primary session is attempting to enter the network, it will try to find a feasible scheduling solution based on unused time slots along its path (without altering the current scheduling for the other active primary sessions). If such a feasible solution does not exist for a new session, it means that the network cannot support this new primary session and it has to be dropped (lost).

For the secondary network $\mathcal{S}$, suppose its new session arrivals also follow a Poisson process. Each new secondary session consists of a source and a destination node and employs shortest path (unicast) routing. For IC, we assume each secondary node is equipped with multiple antennas (as in Fig. 4.1). Suppose each new secondary session has a rate requirement, which corresponds to a number of data streams in MIMO. To enter the network, the new session must ensure that in each time slot along its path: (i) its interference to the primary receivers is canceled; and (ii) the interference from the primary transmitters to the secondary receivers is canceled. The interference in (i) and (ii) is known as inter-network interference. In addition, the new session must also take care of potential mutual interference and self interference within its own secondary network (also known as intra-network interference). Only if the new session can take care of both inter- and intra-network interference successfully can it be admitted into the network. If the new secondary session can be supported (underlay coexistence), its holding time will follow certain distribution. Upon completing its hold time, the secondary session terminates and leaves the network. If the new session cannot be supported for any reason, it has to be dropped (lost).

The above network setting and session behavior reflect the dynamic traffic patterns of primary and secondary sessions in an operational environment. The goal of this chapter is to develop a fast online algorithm for the secondary network to handle such traffic dynamics. In particular, we want our online algorithm to meet the following objectives:

- When a new secondary session initiates, the algorithm must be able to make fast decision on whether or not it can join the network. Such decision must be made through distributed computation based on information stored locally at the nodes along the path of the new

session. An "admit" decision for a new secondary session must successfully address inter- and intra-network interference that is required for underlay coexistence.

- When a new primary session enters the network, existing secondary sessions must make a quick assessment on its impact and formulate a plan to accommodate this new primary session. This include allocation of addition DoFs (if available) for IC. In the extreme case, one (or more) secondary sessions may need to exit the network as the primary session always have pre-emptive priority in terms of spectrum access.

- At all time, our online algorithm must ensure that IC is feasible at the PHY layer for all MIMO transmitters and receivers. By feasible at the PHY layer, we mean that there exist a feasible set of precoding vectors at the secondary transmitters and a feasible set of decoding vectors at the secondary receivers at the PHY layer so that all data (in both primary and secondary networks) can be transported free of interference.

- For performance, we hope our online distributed algorithm can offer a competitive performance when compared to an offline centralized algorithm. Although the latter is not practical for implementation in an online dynamic network environment, it offers a benchmark for comparison and can be used to measure the quality of our online distributed algorithm.

## 4.4 An Online Algorithm

In this section, we present our design of an online distributed algorithm to handle dynamic arrival/departure of the primary and secondary sessions in the underlay coexistence paradigm. The crux of the algorithm is distributed resource allocation (DoFs on the secondary nodes for SM and IC) and the use of local information to accomplish traffic management. With dynamic traffic arrival/departure, the online algorithm must achieve underlay coexistence at all time, i.e., the primary nodes do not feel the presence or activities of the secondary nodes. There are four types of events that constitutes traffic dynamics: initiation of a new secondary session, termination of an existing secondary session, initiation of a new primary session, termination of an existing primary session.

Among these four types of events, the initiation of a new secondary or primary session needs most considerations. When a new secondary session initiates in the network, the online algorithm should make a link-by-link based decision on whether or not at each node along the path there are enough DoFs (over $T$ time slots) to support SM and intra/inter-network IC. When a new primary session arrives, the secondary nodes must take immediate actions to ensure that they will not interfere with the new primary session. Since our algorithm is online and distributed in nature, many race conditions (possible concurrent events) must be addressed. Finally, we must ensure that the DoF allocations at the secondary nodes for SM and IC are indeed feasible at the PHY layer at all time. That is, we must guarantee that one can come up with feasible precoding/decoding vectors at each secondary node to support the proposed DoF allocations.

In this section, we present a distributed algorithm to address the above problems. In Section 4.4.1, we define the set of local information that needs to be maintained at each secondary node. In Sections 4.4.2 to 4.4.5, we present the details of our algorithm to handle the four types of traffic dynamics, with emphasis on new secondary and primary session arrivals. In Section 4.4.6, we show how to solve different race conditions that may occur.

### 4.4.1 Data Structure at Secondary Node

Recall that the secondary nodes have full responsibility in canceling interference to/from the primary network to achieve underlay coexistence. This is a very challenging objective for an online distributed algorithm, particularly when the primary network is not required to communicate directly with the secondary network in the underlay paradigm. To address these challenges, we make the following assumptions and provide necessary justifications.

- *(i) Network topology.* We assume the primary network and the secondary network are each fully connected on its own. That is, any primary node can reach another primary node via single or mult-hop of primary relay nodes. The same also holds true for any secondary node.

- *(ii) Node location information.* We assume that each secondary node has precise information about its location. This can be made possible by the wide availability of GPS capabilities in mobile devices.

- *(iii) Eavesdropping.* We assume the secondary nodes can listen to all communications among the primary nodes. This is important for the secondary nodes to sense the activities of the primary sessions and their transmission/reception behaviors in each time slot.

- *(iv) Control channel.* We assume there is a separate control channel available for the secondary nodes to exchange control information. Control information for the secondary network may propagate one or more hops to reach other secondary nodes..

- *(v) Primary session activity.* For flow (session) level traffic management, we assume there is an explicit link-by-link initiation (set-up) and termination (tear down) phase for each primary session. This assumption will allow DoF allocation (configuration of precoding/decoding vectors) on the secondary nodes to be performed on a feasible time scale.

Among the five assumptions, the eavesdropping assumption is the strongest. The goal of this assumption is to have at least one secondary node to overhear the transmission of each primary node. This assumption is necessary in the development of our online distributed algorithm. Based on these assumptions, some important issues can be addressed. For example, the location of each primary node can be derived, through many available methods in the literature (e.g., [41, 73]).

We now describe the set of local information that needs to be maintained at each secondary node. Our online distributed algorithm will use this local information to make flow management decisions. At each secondary node $i$, we maintain the following information:

- $\lambda_i^{\mathrm{SM}}(t)$ and $\lambda_i^{\mathrm{RM}}(t)$: $\lambda_i^{\mathrm{SM}}(t)$ is the number DoFs used for SM (either as a transmitter or a receiver) at node $i$ in time slot $t$. $\lambda_i^{\mathrm{RM}}(t)$ is the remaining available DoFs at node $i$.

- $\mathcal{X}_i(t)$ and $\mathcal{Y}_i(t)$: These two sets are used to handle inter-network interference to/from the primary nodes. $\mathcal{X}_i(t)$ is the set of node $i$'s neighboring primary transmitters that are active

in time slot $t$, while $\mathcal{Y}_i(t)$ is the set of $i$'s neighboring primary receivers that are active in time slot $t$. Based on our assumption, the secondary nodes can overhear the primary nodes activities, including all control messages. Together with the derived location information of the primary nodes, the secondary nodes can deduce the set of primary nodes that fall in $\mathcal{X}_i(t)$ and $\mathcal{Y}_i(t)$.

- $\alpha_i^j(t), \beta_i^j(t)$ and $\eta_i^j(t)$: These variables are used to handle intra-network interference among the secondary nodes. $\alpha_i^j(t)$ is the number of DoFs being transmitted in time slot $t$ by a secondary transmitter $j$ that is a neighboring node of $i$. $\beta_i^j(t)$ is the number of DoFs being received in time slot $t$ by a secondary receiver $j$ that is a neighboring node of $i$. $\eta_i^j(t)$ is a binary indicator (0 or 1) to denote whether node $i$ is responsible for IC to/from secondary node $j$ in time slot $t$. $\alpha_i^j(t)$ and $\beta_i^j(t)$ are relatively easy to obtain under our five assumptions.

- Channel state information (CSI): The secondary nodes need to have CSI to perform IC (to/from the primary nodes and within the secondary nodes). To estimate CSI between a secondary nodes and its neighboring primary nodes, there are two scenarios. First, if the signal from the primary node can be successfully decoded at the secondary node, then the secondary node can estimate CIS by comparing the decoded signal and the actually received one. On the other hand, if the signal from the primary node cannot be successfully decoded at this secondary node, then based on Assumption *(iii)*, there is another secondary node that is in the neighborhood of the primary node can hear and decode the same signal and broadcast this information to other secondary nodes. Again, by comparing the received (but unable to decode) copy of the signal and the successfully decoded copy of the same signal, the secondary node can estimate the CSI. For either case, based on the reciprocity property of a wireless channel [62], we can derive the CSI in the reverse direction as well. To control the overhead of CSI, we can limit such estimate only during the period when the primary nodes are active and perform such estimates periodically (instead of continuously). The estimation of CSI within the secondary nodes are much easier as it is independent of the primary nodes. Given that the secondary nodes can share control information, we could employ a commonly known pilot signal sequence at a secondary transmitter for CSI estimation. The

neighboring secondary receivers can compare the received copy of the pilot signal sequence with its known version and derive the CSI.

### 4.4.2 Initiation of a New Secondary Session

We first consider how to handle a new secondary session attempting to enter the network. As discussed, the routing path can be found by standard ad hoc routing protocol (e.g., AODV). Denote $f$ as this new session and its source and destination node as $s_f$ and $d_f$, respectively. Suppose that the new session wants to send $R$ data streams from $s_f$ to $d_f$. We assume the the number of antennas at each node is $A$. For each node $i$ along the path, it stores the previous node ($i.prev$) and next node ($i.next$) information along the path.

To determine whether the new secondary session can be supported while achieving underlay coexistence, we perform hop-by-hop examination/update on each link (more precisely, the two nodes of each link) along the path. We denote the Tx and Rx as the transmit and receive nodes of this link, respectively. We start with the first link. Given that there are $T$ time slots in a frame, we begin with the first time slot ($t = 1$).

- For inter-network IC, the Tx node of this link must use its available DoFs to cancel all interference in $\mathcal{Y}_{\text{Tx}}(t)$. Likewise, the Rx node of this link must use its available DoFs to cancel all interference from $\mathcal{X}_{\text{Rx}}(t)$.

- For intra-network IC, the Tx node of this link must use its (remaining) available DoFs to cancel its interference to all active secondary receivers in time slot $t$, which is $\beta^j_{\text{Tx}}(t)$ for each neighboring receive node $j$. Likewise, the Rx node of this link must use its (remaining) available DoFs to cancel the interference from all active neighboring transmit nodes in time slot $t$, which is $\alpha^k_{\text{Rx}}(t)$ for each neighboring transmit node $k$.

- After DoFs are allocated at this link (on Tx and Rx) for inter- and intra-network IC, we check how many (remaining) DoFs are available at Tx and Rx. If both nodes have at least $R$ DoFs available, then all $R$ data streams can be supported in this time slot; otherwise, we move on

to the next time slot, until all $R$ data stream can be supported (possibly over multiple time slots, e.g., one data stream in time slot 1, one data stream in time slot 3, etc.) or we conclude that the $R$ data streams cannot be supported on this link over all time slots.

If the first link can accommodate $R$ data stream for this new secondary session $f$ over $T$ time slots, then the Tx and Rx nodes of this link send the proposed new scheduling information (i.e., $\lambda_{\mathrm{Tx}}^{\mathrm{SM}}(t)$ and $\lambda_{\mathrm{Rx}}^{\mathrm{SM}}(t)$) and their transmission status (i.e., transmitter or receiver) to the transmit and receive nodes of the second link. Both the transmit and receive nodes of this link can obtain the new proposed scheduling information for Tx and Rx. Now we are done with the first link and can move on to the second link. Then both the transmit and receive nodes of the second link first update their maintained information $\alpha$ and $\beta$ based on the message they received, and then follow the same process as the first link. If the second link can accommodate $R$ data stream for this new secondary session $f$ over $T$ time slots, then the Tx and Rx nodes of this second link send their proposed new scheduling information (i.e., $\lambda_{\mathrm{Tx}}^{\mathrm{SM}}(t)$ and $\lambda_{\mathrm{Rx}}^{\mathrm{SM}}(t)$) and their transmission status together with the first link's information on each time slot to the transmit and receive nodes of the next link. Note that there is no need to propagate this new scheduling information to upstream nodes as the scheduling decisions there have already been completed. The link-by-link scheduling process continues until either it is successful for all links or unsuccessful at some link. In the event of end-to-end successful scheduling, the destination secondary node will broadcast scheduling and resource allocation information on behalf of all nodes on the route to other nodes in the secondary network (in the dedicated control channel). The neighboring secondary nodes will update $\alpha$ and $\beta$ in their local information upon receiving the broadcast information. After broadcast, the destination node will return a positive ACK message toward its source indicating that underlay coexistence is achievable along the entire path. Upon receiving this positive ACK in the reverse direction, each node along the path will configure its precoding and decoding vectors at the PHY layer based on the proposed DoF allocation for SM and IC. When the source node receives this positive ACK, it can start transmitting $R$ data streams. On the other hand, if any link fails to support $R$ data streams over its $T$ time slots, then the Tx node of that link will generate a negative ACK message and send it in the reverse direction toward the source. Each upstream node along

the reversed route will discard the proposed DoF allocation and erase any proposed updates. Each node along the reverse path will simple continue its current operation without making any updates on the control plane. Upon receiving the negative ACK at the source, the source node will drop the new incoming secondary session (lost).

**Overhead Analysis and Computation Complexity** For overhead, we count the total number of control messages involved in the process. Recall that all control messages in the secondary network are supported on a separate control channel without any interference to the primary network. When a link on the path can support the $R$ data streams over $T$ time slots, the transmit and receive nodes of this link will each generate a message and pass on to the next link. This requires 2 messages. Since the number of links along the rout is no more than $(S-1)$, the number of such control message is no more than $2(S-1)$. When the admission test is successful at the last link, the destination node will broadcast a message containing each node's scheduling information to all nodes in the secondary network. The number of messages involved in this broadcast is no more than $S$. The destination node also sends a positive ACK on the reverse path toward the source node, which requires relaying this message by at most $S-1$ times. So the total control messages is $O(2(S-1)) + O(S) + O(S-1) = O(S)$.

For each node along the path in a time slot, its total number of DoFs is $A$. So the number of allocations of DoFs for IC and SM is no more than $A$ times. Since there is a total of $T$ time slots in a frame, the complexity at each node is $O(TA)$. Since the number of nodes on the path is no more than $S$, the total computational complexity is $O(TAS)$.

### 4.4.3 Termination of a Secondary Session

When a secondary session $f$ decides to terminate, it can cease to transmit data stream immediately on the data plane. On the control plane, an explicit link-by-link tear-down process is needed to release DoFs used for IC and SM. We start from the first link. Both the transmit node Tx and receive node Rx of this link will send their updated transmission information for this session (i.e., how the $R$ data streams are removed over $T$ time slots) to the transmit and receive nodes of the next

link along the path. This is done in the control channel. This information will eventually propagate link-by-link toward the destination. The Tx and Rx nodes of the first link will check $T$ time slots and release the DoF allocation for SM and IC on this link, and update their scheduling and resource allocation information in corresponding time slots. For example, the source node will release its DoF allocation for SM and IC in each time slot $t$ that is used to support the $R$ data streams, update its $\lambda_{\mathrm{Tx}}^{\mathrm{SM}}(t)$ and $\lambda_{\mathrm{Rx}}^{\mathrm{RM}}(t)$, and set binary variable $\eta_{\mathrm{Tx}}^{j} = 0$ for IC to each neighboring receive node $j$. The receive node Rx will release its DoF allocation for SM and IC in the time slot $t$, update its scheduling $\lambda_{\mathrm{Rx}}^{\mathrm{SM}}(t)$ and $\lambda_{\mathrm{Rx}}^{\mathrm{RM}}(t)$, and set binary variable $\eta_{\mathrm{Rx}}^{k} = 0$ for IC for each neighboring transmit node $k$. Note that the release of DoFs at a transmit or receive node corresponds to freeing up the variables in the precoding or decoding vectors for SM and IC at the node. Given the variables are freed up here, it is always feasible at the PHY layer. Once these updates are completed for the first link, we move on to the second link. The transmit and receive node of the second link will send their updated transmission status in each time slot, along with the information received from the first link, to the next link (third link). Both the transmit node Tx and receive node Rx of the second link will send their transmission information for this session (i.e., how the $R$ data streams are supported over $T$ time slots) to the transmit and receive nodes of the next link along the path. The Tx and Rx nodes of the second link will then release the DoFs allocated for SM and IC for the session on this link, and update their scheduling and resource allocation information in corresponding time slots. Once we are done with the second link, we move on to the third link and so forth. This process continues until reaching the destination node. The destination node has the aggregated transmission information for this session from all nodes on this route. It broadcasts this information to other all in the secondary network (in the dedicated control channel), announcing the termination of this session. Note that the termination operation is done one way from source to destination (in contrast to a round trip in session initiation). The neighboring nodes that previously used DoFs to cancel interference to/from the terminated session will update $\alpha$, $\beta$, and $\eta$ in their local information upon receiving the broadcasted message, and release the DoFs for IC to those nodes on the terminated session.

**Overhead Analysis and Computational Complexity** When a link is torn down, the transmit and

receive nodes of this link will each generate a message and pass on to the next link. This requires 2 messages. Since the number of links along the rout is no more than $(S-1)$, the number of such control message is no more than $2(S-1)$. In the end, the destination node will broadcast a message containing each node's transmission information to all nodes in the secondary network. The number of messages involved in this broadcast is no more than $S$. So the total control messages is $O(2(S-1)) + O(S) = O(S)$.

For the termination of a secondary session, all secondary nodes along the path and those secondary nodes that have IC relationship with this session will need to update DoF scheduling information on relevant time slots. Since there are $T$ time slots and at most $S$ secondary nodes, the complexity is $O(TS)$.

### 4.4.4 Initiation of a New Primary Session

We now consider how to handle the network scenario where a primary session initiates in the network. The primary node can use whatever routing protocol it prefers to find a route. In underlay coexistence paradigm, the primary nodes do not notice the activities of the secondary nodes. They only need to be concerned with other active primary nodes in the network. The primary nodes can use whatever scheduling algorithm to decide whether the new session can be admitted. Since we do not mandate a specific scheduling algorithm for the primary nodes, the discussion of scheduling algorithm for the primary nodes is beyond the scope of this chapter. We are only interested in how the secondary nodes respond when a new primary session initiates so as not to interfere with any of the primary nodes (underlay).

The main technical challenge here is that, in a time slot, how a secondary transmit node can cease its transmission when a primary node starts to transmit in the same time slot? This is a fundamental problem in spectrum sharing. In the context of underlay coexistence, we propose the following solution. We divide each time slot for a secondary node into two parts: a small interval (on the order of several bits) for spectrum sensing and the remaining part for actual transmission [25]. During the spectrum sensing interval, if the secondary transmitter find that there

is change in neighboring primary transmitter's scheduling behavior (e.g., becoming active in this time slot), then the secondary node cease to transmit in the remaining interval in this time slot. Based on our eavesdropping assumption, the secondary node can listen and decode the control information (in the packet header) of the primary transmitter. It will broadcast the activation of this new primary session to all other secondary nodes in the network (in the dedicated control channel for the secondary network). Given that the primary session has multiple nodes along its path, all neighboring secondary nodes will need to broadcast the change of scheduling behavior of the primary nodes along the path. This may incur considerable overhead in the number of control messages. So some aggregation of control messages at the secondary nodes is necessary. We will discuss the the complexity of this operation shortly.

Upon hearing the activation of a new primary session, all secondary nodes that have interference with the primary session will immediately freeze their transmissions. They will also notify the source nodes of these involved secondary sessions (on control channel), who will immediately suspend transmission for these sessions. Upon hearing that the primary session is successfully admitted into the primary network, the neighboring secondary nodes will update their local information for $\mathcal{X}$ and $\mathcal{Y}$, based on the new scheduling behavior at the primary nodes. If the new primary session cannot be admitted into the primary network, then there is an explicit negative ACK message returning to the source node. Upon hearing this negative ACK message, the secondary nodes that have frozen their transmissions will generate a RESUME message back toward their source nodes so that those suspended secondary sessions can resume their transmissions.

After the new primary session is admitted into the network, those secondary sessions that are impacted by the new primary session will need to go through a re-admission process. The re-admission process for each session is the same as that in Section 4.4.2, except that we need to address the race condition of multiple such secondary sessions. In Section 4.4.6, we employ token passing to solve the race condition so that competing secondary sessions are handled one at a time. Such sequential handling of re-admission processes of concurrent secondary sessions is critical to achieve IC feasibility at the PHY layer. After going through a re-admission process, the impacted secondary session can either be admitted to re-enter the network or be terminated (due to lack of

resources on the path).

**Overhead Analysis and Computation Complexity** When a new primary session is admitted, the neighboring secondary nodes will broadcast the new scheduling information of the primary nodes to all other secondary nodes in the network. The total number of messages involved in these broadcasts is no more than $O(S^2)$. Note that this is the worst case overhead. In practice, we can aggregate multiple incoming control messages and self-generated control message at a secondary node into a single broadcast. The lower bound is $\Omega(S)$.

For those secondary sessions that are impacted by the new primary session, the overhead for each of them to go through a new admission process is the same as that in Section 4.4.2, which is $O(S)$. Since there $F$ sessions, the overhead is $O(FS)$.

After a new primary session is admitted, a neighboring secondary node will update $\mathcal{X}$ and $\mathcal{Y}$ on each time slot, based on the new scheduling information from the primary nodes. Since there is a total of total $S$ nodes, the complexity is $O(TS)$.

For each secondary session that is impacted by the new primary session, the complexity is the same as that for the initiation of a new secondary session, which is $O(TAS)$. Since there are at most $F$ sessions to be impacted, the complexity for re-admission process is $O(FTAS)$.

### 4.4.5 Termination of a Primary Session

Based on assumption *(v)*, the termination of a primary session employs an explicit link-by-link tear-down process. Upon hearing this control message along the path of a primary session, the neighboring secondary nodes will broadcast the change of scheduling behavior of the primary nodes along the path.

Note that the termination of a primary session will not affect the current transmission behavior of active secondary sessions. Each secondary node can still use its current scheduling for its own transmission (SM). But the IC responsibilities on the neighboring secondary nodes will change. Upon receiving the broadcast messages, a secondary node will release the DoF allocation (free-

ing up the variables in the precoding/decoding vectors) for IC to/from the primary nodes on the terminated primary session, and update their locally maintained information $\mathcal{X}$ and $\mathcal{Y}$.

**Overhead Analysis and Computation Complexity** Upon hearing the termination of a primary session, each neighboring secondary node broadcasts the change of the primary nodes' scheduling behaviors. The total number of messages involved in these broadcasts is between $\Omega(S)$ and $O(S^2)$.

Each neighboring secondary node along the path of the terminated primary session needs to update their local information $\mathcal{X}$ and $\mathcal{Y}$. This update needs to be done on each time slot. So the worst case complexity is is $O(TS)$.

### 4.4.6 Coping the Race Problem

A major challenge in our design of online distributed algorithm is to address race condition. For example, the processing of a new secondary session arrival may take one round trip time to travel across network diameter. During this time, another new secondary session arrival may also occur. Since the IC responsibilities on the nodes in the latter session may depend on the first session, a blind processing of the latter session concurrently with the first one may result in infeasible DoF allocation at the PHY layer.

There are two approaches to address such race condition, both employs token passing. The first approach is similar to token ring, where a token is passed cyclically among the secondary nodes. A new secondary session is allowed to start its link-by-link DoF test only when the token is passed to the source node of the session. Once the source node holds the token, the corresponding session is the only new session that is under link-by-link DoF examination. Upon its completion, the source node will pass the token to the next node in the cycle and so forth. The advantage of this approach is that it is fully distributed. A lost token may be recovered through timeout. But the disadvantage is that the cycle time (for a token to travel around all secondary nodes) may be long $O(S)$.

To speed up token passing time, the second approach employs a dedicated secondary node to serve as a token controller. This can be done through the distributed *leader election* algorithm [39],

which has a message overhead of $O(S\log(S))$ and only needs to be done once. Each secondary node will need to maintain a route from itself toward the token controller. When a new secondary session arrives, its source node will send a token request to the token controller node, requesting for a token. The token controller will grant a token only if it currently holds the token (not being taken by another secondary source node). Otherwise, the new token request will be queued until the token returns to the token controller. This token passing approach will effectively handle secondary-secondary race condition. Although this approach relies on a dedicated secondary node (as token controller), it offers faster passing among the secondary nodes. To cope single point failure, another secondary node may be used as a back up token controller (similar to DNS infrastructure). We adopt this approach to resolve secondary-secondary race condition.

Note that there is no race condition when a secondary session leaves the network (termination). When a secondary session decides to terminate, it can cease data transmission immediately. As discussed in Section 4.4.3, the session tear-down process is done on a link-by-link basis by releasing DoF allocation for SM and IC. This process continues until reaching the destination node. To minimize control overhead, only the destination node broadcasts the tear-down of the path (and all nodes involved) to other nodes in the secondary network. Upon receiving this tear-down broadcast, those relevant nodes can release their DoF allocation for IC to these nodes on the session route. Since reconfiguring precoding/decoding vectors at a secondary node to release DoFs is guaranteed to be feasible, any concurrent operation involving a secondary session's departure is not considered a race condition.

When a new primary session initiates, a race condition may occur when a new secondary session also arrives. This is easy to handle as we assume the secondary nodes can eavesdrop the control channel (in band or out band) of the primary network. So upon identifying a new primary session's initiation, any new secondary session initiation activity will freeze until the new primary session is processed.

## 4.5 Physical Layer Feasibility

In Sections 4.4.2 to 4.4.5, we have taken every step to ensure underlay coexistence for the secondary sessions under various traffic dynamics. In this section, we show that the PHY layer feasibility is maintained at each secondary node at all time. By PHY layer feasibility, we mean that there exist feasible precoding/decoding vectors at each secondary node to implement the desired DoF allocation for SM and IC.

**A secondary session initiates** When a new secondary session is admitted into the network, we perform the link-by-link operation to allocate DoF for SM and IC. To achieve underlay coexistence, the nodes on the new secondary session must perform inter-network and intra-network IC. Note that the IC responsibilities on the existing secondary sessions do not change. For nodes along the path of the new secondary session, we start with the first link (containing source node) and work our way toward the last link (containing the destination node). For each node, its DoF allocation for IC follows a sequential order from the source node to the current node. That is, IC to/from those nodes that are after this node along the path of the new secondary session is not the responsibility of this node. Such interference will be taken care of when we consider those nodes later. This sequential accounting of IC responsibility is the basis of our construction of precoding/decoding vectors at each node along the path (from source node toward destination node).

**Theorem 2.** *After a new secondary session is successfully admitted into the network, there exists a set of feasible precoding/decoding vectors at each secondary node along the path based on the DoF allocation for SM and IC in the admission process.*

*Proof.* Our construction of precoding/decoding vectors at each node starts from the source. For the first link, denote $\mathrm{Tx}$ as the source node and $\mathrm{Rx}$ as the receive node. Based on the DoF allocation for SM and IC in time slot $t$ in the admission process, we now show that we can construct a feasible set of precoding vectors at Tx in the same time slot. At node Tx, the local information $\mathcal{Y}_{\mathrm{Tx}}(t)$ contains the neighboring primary receivers in time slot $t$, while $\beta_{\mathrm{Tx}}^{j}$ is the number of data streams being received at neighboring secondary receivers $j$. The secondary node Tx needs to construct the

precoding vectors to cancel all interference to these receivers. Denote the set of these neighboring secondary receivers as $\mathcal{B}$. Suppose that Tx transmits $z_{(\mathrm{Tx,Rx})}$ data streams to Rx in time slot $t$. Denote $\mathbf{u}_{\mathrm{Tx}}^{a}$ as an $A \times 1$ transmit vector at Tx for each data stream $a$ ($1 \leq a \leq z_{(\mathrm{Tx,Rx})}$), and $\mathbf{v}_{\mathrm{Rx}}^{a}$ as an $A \times 1$ receive vector at Rx to receive data stream $a$.

Denote $\mathbf{H}_{(\mathrm{Tx},j)}$ as the $A \times A$ channel matrix between nodes Tx and $j$ ($j \in \mathcal{B}$), and denote $\mathbf{H}_{(\mathrm{Tx},k)}$ as the $A \times 1$ matrix between Tx and the primary receive node $k$ ($k \in \mathcal{Y}_{\mathrm{Tx}}(t)$). We assume all channels $\mathbf{H}_{(\mathrm{Tx},j)}$ and $\mathbf{H}_{(\mathrm{Tx},k)}$ are full ranks. To transmit $z_{(\mathrm{Tx,Rx})}$ data streams from node $\mathrm{Tx}$ to $\mathrm{Rx}$ while achieving underlay coexistence, transmit node $\mathrm{Tx}$ must cancel its interference to neighboring primary receivers in $\mathcal{Y}_{\mathrm{Tx}}(t)$ and neighboring secondary receivers in $\mathcal{B}$. Then, we should have the following constraints:

$$(\mathbf{u}_{\mathrm{Tx}}^{a})^{T}\mathbf{H}_{(\mathrm{Tx,Rx})}\mathbf{v}_{\mathrm{Rx}}^{a} = 1\ , \qquad (1 \leq a \leq z_{(\mathrm{Tx,Rx})}), \tag{4.5.1}$$

$$(\mathbf{u}_{\mathrm{Tx}}^{a})^{T}\mathbf{H}_{(\mathrm{Tx,Rx})}\mathbf{v}_{\mathrm{Rx}}^{b} = 0\ , \qquad (1 \leq a \leq z_{(\mathrm{Tx,Rx})}, 1 \leq b \leq z_{(\mathrm{Tx,Rx})}, a \neq b), \tag{4.5.2}$$

$$(\mathbf{u}_{\mathrm{Tx}}^{a})^{T}\mathbf{H}_{(\mathrm{Tx},k)}v_{k} = 0\ , \qquad (1 \leq a \leq z_{(\mathrm{Tx,Rx})}, k \in \mathcal{Y}_{\mathrm{Tx}}(t)), \tag{4.5.3}$$

$$(\mathbf{u}_{\mathrm{tx}}^{a})^{T}\mathbf{H}_{(\mathrm{Tx},j)}\mathbf{v}_{j}^{q} = 0\ , \qquad (1 \leq a \leq z_{(\mathrm{Tx,Rx})}, j \in \mathcal{B}, 1 \leq q \leq z_{(i,j)}), \tag{4.5.4}$$

where $i$ is the transmit node which transports $z_{(i,j)}$ data streams to secondary receive node $j$. Since each primary receiver has only a single antenna and can only receive one data stream, $v_k$ is a constant for each $k \in \mathcal{Y}_{\mathrm{Tx}}(t)$.

The number of constraints in (4.5.1) and (4.5.2) is $(z_{(\mathrm{Tx,Rx})})^2$. The number of constraints in (4.5.3) is $z_{(\mathrm{Tx,Rx})} \cdot \sum_{k \in \mathcal{Y}_{\mathrm{Tx}}(t)} 1$. The number of constraints in (4.5.4) is $z_{(\mathrm{Tx,Rx})} \cdot \sum_{j \in \mathcal{B}} z_{(i,j)}$. So the total number of constraints is $(z_{(\mathrm{Tx,Rx})})^2 + z_{(\mathrm{Tx,Rx})} \cdot \sum_{k \in \mathcal{Y}_{\mathrm{Tx}}(t)} 1 + z_{(\mathrm{Tx,Rx})} \cdot \sum_{j \in \mathcal{B}} z_{(i,j)} = z_{(\mathrm{Tx,Rx})} \cdot (z_{(\mathrm{Tx,Rx})} + \sum_{k \in \mathcal{Y}_{\mathrm{Tx}}(t)} 1 + \sum_{j \in \mathcal{B}} z_{(i,j)})$. In our DoF allocation at Tx, the total number of DoFs allocated for SM and IC cannot exceed $A$, i.e., $(z_{(\mathrm{Tx,Rx})} + \sum_{k \in \mathcal{Y}_{\mathrm{Tx}}(t)} 1 + \sum_{j \in \mathcal{B}} z_{(i,j)}) \leq A$, where $z_{(\mathrm{Tx,Rx})}$ is the number of DoFs for SM, $\sum_{k \in \mathcal{Y}_{\mathrm{Tx}}(t)} 1$ is the number of DoFs for IC to primary receivers, and $\sum_{j \in \mathcal{B}} z_{(i,j)}$ is the number of DoFs for IC to neighboring secondary receivers. Therefore, the total number of constraints is no more than $z_{(\mathrm{Tx,Rx})} \cdot A$.

In the above constraints, $v_k$ ($k \in \mathcal{Y}_{\mathrm{Tx}}(t)$) are constants, and $\mathbf{v}_{j}^{q}$ ($j \in \mathcal{B}$) belong to the existing

secondary nodes (not on the path of the new secondary session), which are already configured. For precoding vector $\mathbf{u}_{\mathrm{Tx}}^{a}$ for data stream $a$ ($1 \leq a \leq z_{(\mathrm{Tx,Rx})}$), it is an $A \times 1$ vector. So the total number of variables for $z_{(\mathrm{Tx,Rx})}$ vectors at transmit node Tx is $z_{(\mathrm{Tx,Rx})} \cdot A$, which is no less than the number of constraints. On the other hand, since the channels are of full rank and independent of each other, it can be shown that the constraints in (4.5.1), (4.5.2), (4.5.3), and (4.5.4) are linearly independent with each other [59]. So for any *given* $\mathbf{v}_{\mathrm{Rx}}^{b}$ for $1 \leq b \leq z_{(\mathrm{Tx,Rx})}$, we are guaranteed to construct feasible precoding vectors $\mathbf{u}_{\mathrm{Tx}}^{a}$ ($1 \leq a \leq z_{(\mathrm{Tx,Rx})}$) at Tx.

After we construct feasible precoding vectors at Tx. We can construct the decoding vectors $\mathbf{v}_{\mathrm{Rx}}^{b}$ for $1 \leq b \leq z_{(\mathrm{Tx,Rx})}$ in time slot $t$ based on $\mathbf{u}_{\mathrm{Tx}}^{a}$ following the same argument. Therefore, for the proposed DoF allocation for SM and IC in a time slot for the new secondary session, we can show that there exist precoding vectors at Tx and decoding vectors at Rx. After Tx and Rx are configured, we move on to the next link and use the same approach to construct precoding vectors at the transmit node and the decoding vetors at the receive node for the next link and so forth. In essence, since the number of DoFs that can be allocated for SM and IC is no more the number of antennas (i.e., $A$) at each node in the admission process, the number of constraints is no more than the number of variables. Therefore, we can always construct feasible precoding/decoding vectors at each secondary node along the path. This completes the proof. □

**A primary session initiates** After a new primary session successfully joins the network, the impacted sessions should first cease their transmission and then go through a new admission process again. This operation is the same as the initiation of a new secondary session. The feasibility proof at the PHY layer is the same as that for Theorem 2.

**Termination of a primary or a secondary session.** In either case, the secondary nodes involved in IC only need to release DoFs (i.e., freeing the variables in precoding/decoding vectors), this operation is always feasible at the PHY layer.

## 4.6 Performance Evaluation

In this section, we evaluate the performance of our online distributed algorithm to handle traffic dynamics. We organize our evaluation into three parts. In the first part, we study the performance of our online distributed algorithm in terms of lost secondary sessions. As a benchmark, we compare the performance of our algorithm to that of an offline algorithm. In the second part, we examine whether underlay coexistence holds at all time in the network.

### 4.6.1 Parameter Settings

We consider a 50-node primary network and a 50-node secondary network randomly deployed in an $100 \times 100$ area. The location of the primary and secondary nodes are shown in Fig. 4.3. Each primary node is equipped with a single antenna while each secondary node is equipped with four antennas. Both the primary and secondary networks share the same spectrum bandwidth. For generality, we normalize all units for distance and bandwidth with appropriate dimensions. We assume the transmission range and interference range for both the primary and secondary nodes are 30 and 50, respectively. For scheduling, a time frame is divided into four time slots (i.e., $T = 4$).

We assume the primary and secondary session arrivals each follow a Poisson process. The arrival rate for the primary and secondary sessions will be specified in the respective performance studies. The holding time for each primary or secondary session follows an exponential distribution with a mean of 1 minute. For each primary session, it can only request 1 data stream. But for each secondary session, it can request $R$ data streams. We set $R = 2$ in our study.

### 4.6.2 Lost Secondary Sessions

A key performance measure of our proposed online distributed algorithm is its ability to accommodate as many new secondary sessions into the network as possible while meeting underlay

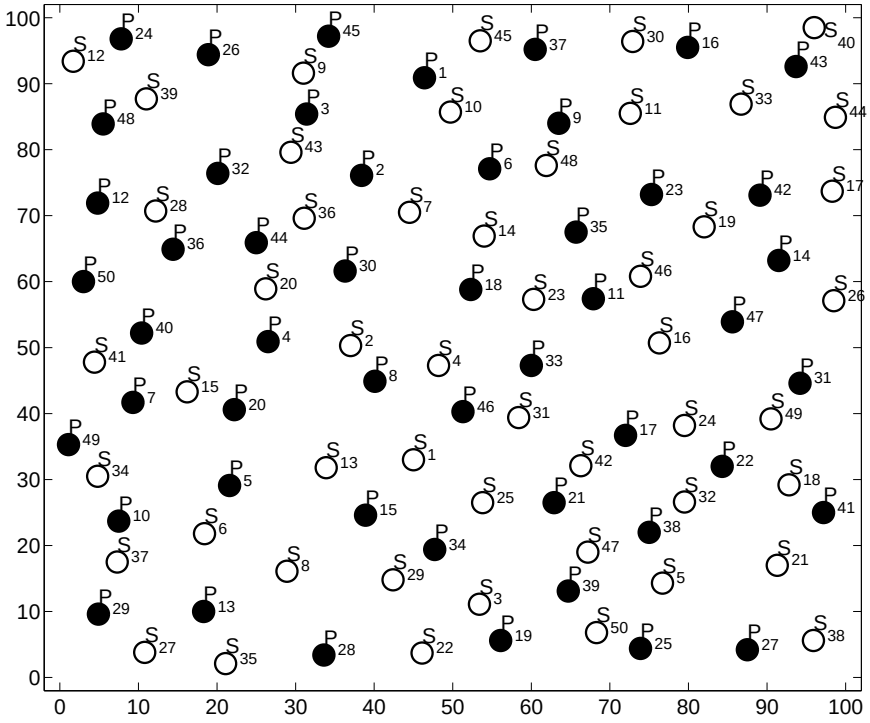



Figure 4.3: The locations of the primary and secondary nodes.

coexistence requirements. A secondary session may be lost under two circumstances: (i) it may be rejected by our algorithm when it initially arrives to the network; (ii) it may be suspended due to the arrival of a new primary session and subsequently cannot be re-admitted into the network. In both case, we consider that secondary session is lost.

To measure the performance of our distributed algorithm, we compare it to that of an offline algorithm. For fairness, an offline algorithm will employ the same shortest path routing as our online algorithm. The difference is that an offline algorithm will perform a global optimization (among all secondary sessions) to find a feasible DoF allocation. Under this new feasible DoF allocation, a secondary node with its current precoding/decoding vectors may need reconfigure these vectors, which is hardly practical in real time. In contrast, for an online algorithm, it will not alter the DoF allocation on those secondary nodes that are already active. It will only allocate DoFs (and configure precoding/decoding vectors) on the nodes that are traversed by the new secondary session.

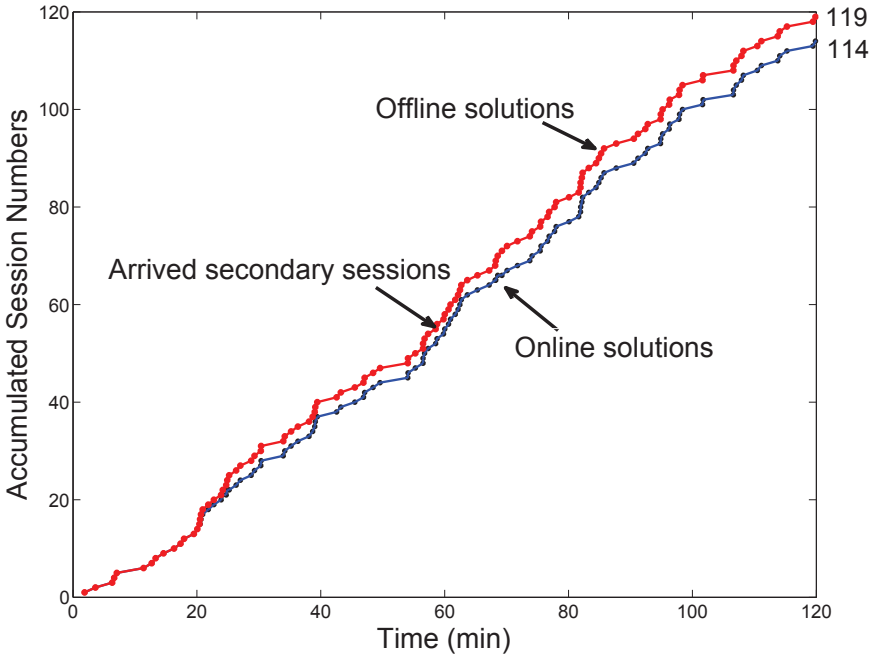



Figure 4.4: Cumulative secondary arrivals, admitted secondary arrivals by offline algorithm, and admitted secondary arrivals by our online algorithm. Both primary and secondary session arrival rates are 1 per minute.

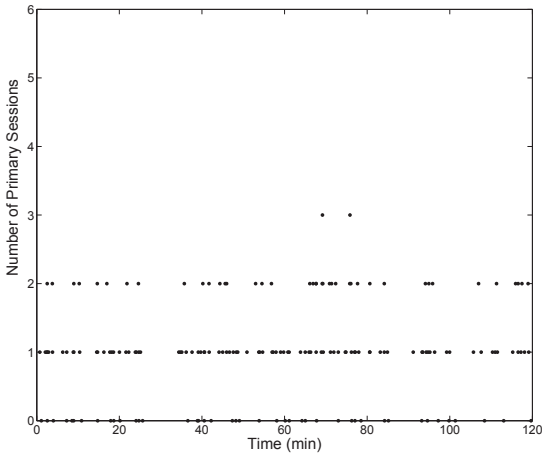
In Chapter 2, a global (centralized) optimization problem formulation is given that an offline algorithm shall solve. It is in the form of mixed integer linear program (MILP), which is NP-hard in general [57]. We use a commercial CPLEX solver for the MILP problem and set the termination time to 1 hour. There are several possibilities: (i) before or by the termination time, CPLEX finds a new feasible DoF allocation for all secondary sessions; (ii) before or by the termination time, CPLEX finds that there does not exist a feasible DoF allocation to accommodate the new session; (iii) by the termination time, CPLEX still cannot find a feasible DoF allocation (due to the complexity of the global optimization problem). Under case (i), the new session is admitted into the network under the offline algorithm, while under cases (ii) and (iii), we consider the offline algorithm cannot accommodate the new secondary session (i.e., lost).

Fig. 4.4 shows the cumulative total arrivals of secondary sessions, admitted arrivals of secondary sessions by the offline algorithm, and admitted arrivals of secondary sessions by our online

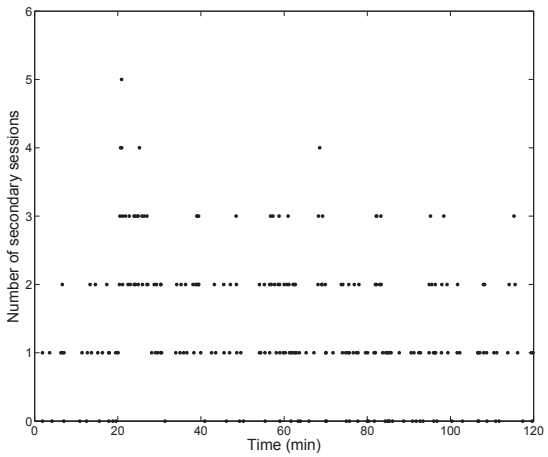
algorithm. Both the primary and secondary sessions' arrival rates are 1 per minute. Note that the curves for cumulative total arrivals of secondary sessions and admitted arrivals of secondary sessions by the offline algorithm coincide completely, indicating that all new secondary sessions are admitted without any loss. This clearly represents operation in low traffic load region. In this region, we find that the online algorithm performs very well. Over a period of 2 hours, there is a total of 119 new secondary session arrivals, all of them can be admitted by the offline algorithm, while our online algorithm can admit 114 (96%).

To show the session dynamics in Fig. 4.4, Fig. 4.5 (a), (b) and (c) show the number of active primary sessions in the network, the number of secondary sessions that can be admitted into the network by the offline algorithm, and the number of secondary sessions that are admitted into the network by our online algorithm, all over a 2-hour period, respectively. We find that the number of primary sessions vary from 0 to 3 while the number of secondary sessions vary from 0 to 4.

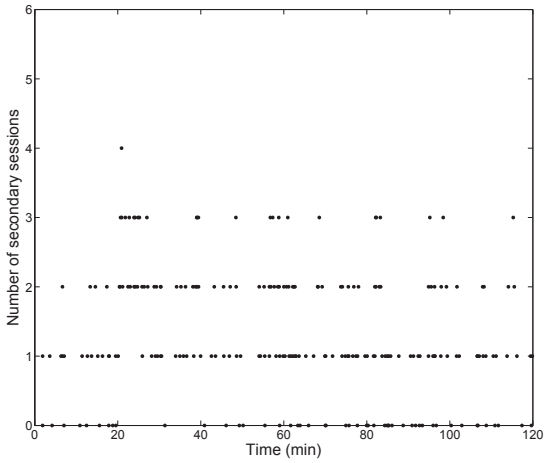
We now increase traffic load on the network by increasing the arrival rate of new secondary sessions. The arrival rate for the primary sessions and session hold time are the same as before. Figures 4.6 and 4.7 show the cumulative secondary arrivals, admitted secondary arrivals by offline algorithm, and admitted secondary arrivals by our online algorithm over a 2-hour period when the secondary session arrival rates are 5 and 10, respectively. Clearly, we find that there is a gap between the curves of cumulative secondary arrivals and admitted secondary arrivals by offline algorithm, indicating that the new secondary arrivals are lost even under the offline algorithm. This gap widens as the arrival rate of new secondary sessions increases from 5 to 10 per minute. We now compare the performance of our online algorithm with that of the offline algorithm. When the secondary session arrival rate is 5 per minute (moderately heavy load), there are 385 new sessions admitted by our algorithm while 444 admitted by the offline algorithm. The ratio between the two is 87%. When the secondary session arrival rate is 10 per minute (heavy load), there are 468 new sessions admitted by our algorithm while 569 admitted by the offline algorithm. The ratio between the two is 82%. Fig. 4.8 shows the ratios between admitted secondary sessions by our online algorithm and that by the offline algorithm under a wide range of traffic load. We find the minimum ratio is 82%, which indicates that our online algorithm is competitive.




(a) The number of active primary sessions in the network.




(b) The number of secondary sessions that can be admitted into the network by the offline algorithm.




(c) The number of secondary sessions that are admitted into the network by our online algorithm.

Figure 4.5: The number of active primary sessions in the network, the number of secondary sessions that can be admitted into the network by the offline algorithm, and the number of secondary sessions that are admitted into the network by our online algorithm, all over a 2-hour period.

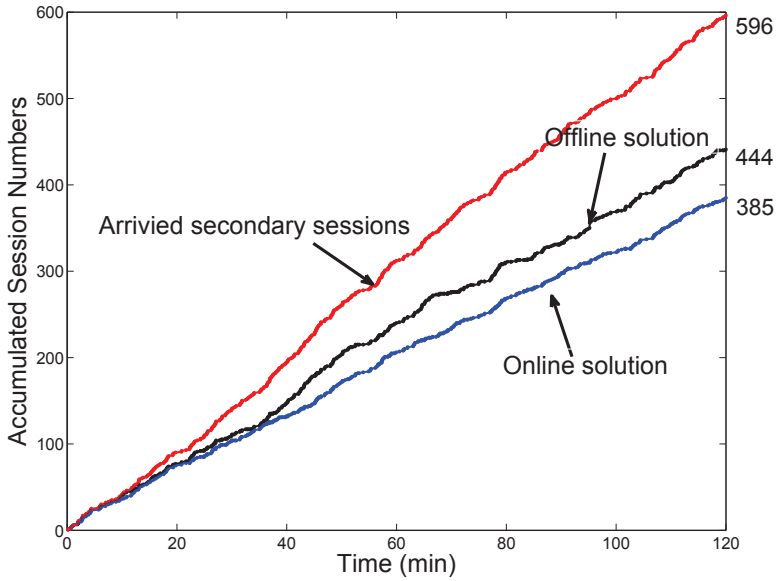



Figure 4.6: Cumulative secondary arrivals, admitted secondary arrivals by offline algorithm, and admitted secondary arrivals by our online algorithm. Primary and secondary session arrival rates are 1 and 5, respectively.

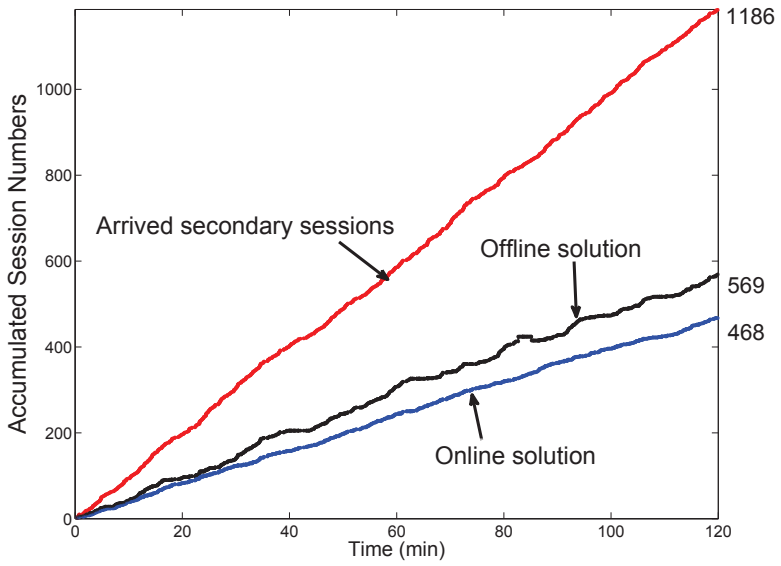



Figure 4.7: Cumulative secondary arrivals, admitted secondary arrivals by offline algorithm, and admitted secondary arrivals by our online algorithm. Primary and secondary session arrival rates are 1 and 10, respectively.

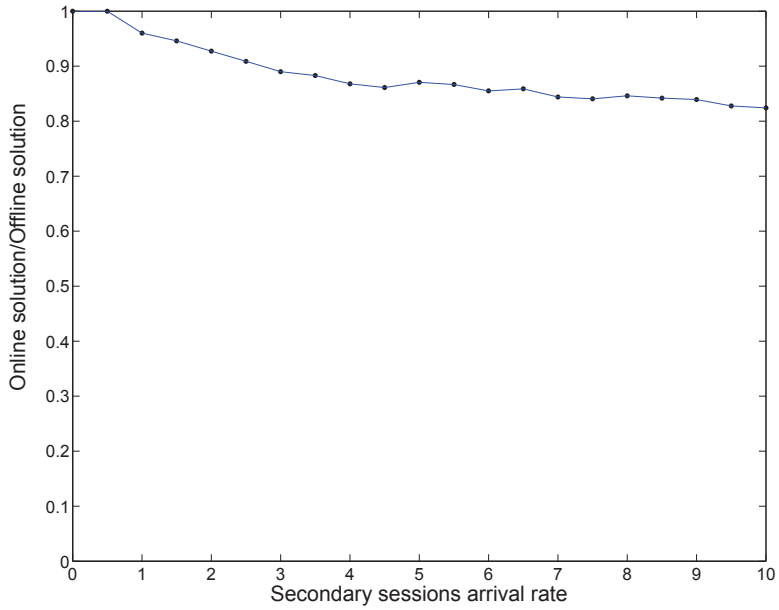



Figure 4.8: Ratios between admitted secondary sessions by our online algorithm and that by the offline algorithm with different secondary sessions arrival rate.

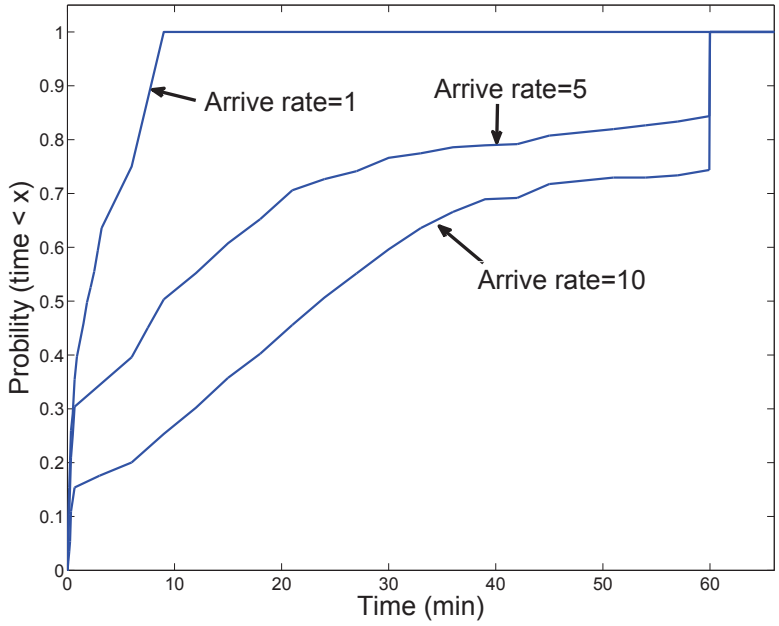



Figure 4.9: The CDFs of computation time by the offline algorithm when the secondary sessions arrival rates are 1, 5, and 10 respectively. The cutoff termination time for the offline algorithm is set to 1 hour.

We now compare the computational time by our online algorithm and that by the offline algorithm. The computation time by our online algorithm includes local computation time at secondary nodes (negligible) and communication time among the secondary nodes. The latter is on the same order of round trip time between any two secondary nodes (source and destination) in the network, which is again very small. On the other hand, the computational time by the offline algorithm is the time used by CPLEX solver, with a cutoff termination time of 1 hour. Fig. 4.9 shows the CDFs of computational time by the offline algorithm when the secondary sessions arrival rates are 1, 5, and 10 respectively, which correspond to our studies in Figs. 4.4, 4.6 and 4.7. Note that even under very light traffic load (with secondary session arrival rate being 1 per minute), more than 20% of new sessions still require at least 5 minutes for the CPLEX solver to find a feasible solution. This is not acceptance for the arrival rate, which is 1 per minute. When the secondary session arrival rate increases, the situation deteriorates. For example, when the secondary session arrival rate is 5 per minute (moderately heavy load), more than 50% of new sessions require at least 5 minutes for the CPLEX solver to find a feasible solution while more than 15% of sessions exceed the cutoff termination time (1 hour). The situation for the case when the secondary session arrival rate is 10 per minute (heavy load) is even worse. The results in Fig. 4.9 shows that even under light load, an offline algorithm is not practical.

### 4.6.3 Validation of Transparent Coexistence

In this section, we examine whether the underlay coexistence of secondary sessions are always maintained by our online distributed algorithm. That is, we want to show that inter-network interference (interference to/from the primary network) and intra-network interference (interference within the secondary network) are all cancelled properly.

For validation, we randomly pick some time instances and examine how interference is canceled. Let's consider time at 17.3 minute in Fig. 4.4 and Figs. 4.5 (a), (b), and (c), when there is a new secondary session arrival ($S_{42} \rightarrow S_{10}$). Fig. 4.10(a) and (b) shows routing and scheduling of primary and secondary sessions before and after the new secondary session arrival. We will

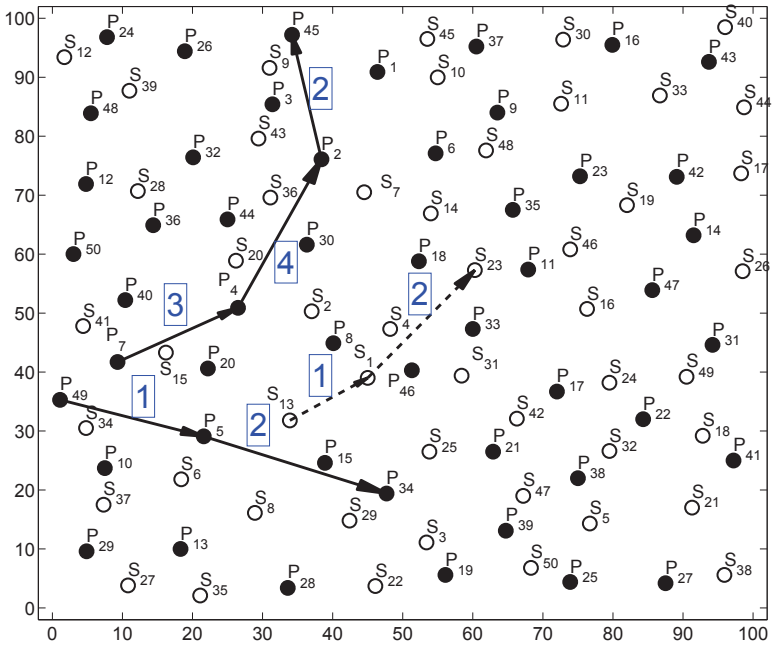
validate whether underlay coexistence holds in each case.

Table 4.1: DoF allocation for SM and IC for the secondary sessions in each time slot before the new session arrives.

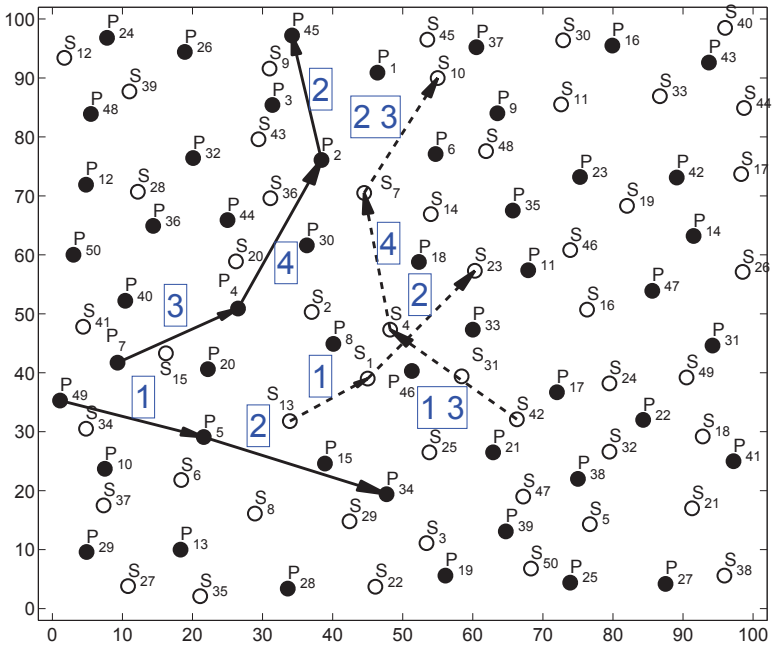
| Time Slot 1 | | | | |
|---|---|---|---|---|
| Node $i$ | TX/RX | DoF for SM | DoF for IC to/from primary nodes | DoF for IC within secondary network |
| $S_{13}$ | TX | 2 | 1 to $P_5$ | NO |
| $S_1$ | RX | 2 | 1 from $P_{49}$ | NO |
| Time Slot 2 | | | | |
| $S_1$ | TX | 2 | 1 to $P_{34}$ | NO |
| $S_{23}$ | RX | 2 | 1 from $P_2$, 1 from $P_5$ | NO |

Before the new secondary session arrives (see Fig. 4.10(a)), there are two primary sessions ($P_7 \rightarrow P_{45}$ and $P_{49} \rightarrow P_{34}$) and one secondary session (i.e., $S_{13} \rightarrow S_{23}$) in the network. The scheduling (in time slot) for each link is marked in a box next to the link. For example, in time slot 1, primary link $P_{49} \rightarrow P_5$ and secondary link $S_{13} \rightarrow S_1$ are active. To illustrate how each interference is canceled, Table 4.1 shows the first two time slots (there is no inter-network interference in time slots 3 and 4 and its discussion is omitted). As shown in Fig. 4.11(a) and Table 4.1, in the first time slot, secondary node $S_{13}$ interferes $P_5$ with 1 DoF. So node $S_{13}$ allocates 1 DoF to cancel this interference. Also, primary node $P_{49}$ interferes $S_1$ with 1 DoF. So node $S_1$ allocates 1 DoF to cancel this interference. Both primary link $P_{49} \rightarrow P_5$ and secondary link $S_{13} \rightarrow S_1$ are active in time slot 1. Since all inter-network interference is canceled by the secondary nodes, underlay coexistence for the secondary nodes holds in time slot 1. In the second time slot, secondary node $S_1$ interferes $P_{34}$ with 1 DoF. So node $S_1$ allocates 1 DoF to cancel this interference. Also primary nodes $P_2$ and $P_5$ interfere with secondary node $S_{23}$, each with 1 DoF. So $S_{23}$ allocated 2 DoF to cancel each of these two interferences. Both primary links $P_5 \rightarrow P_{34}$ and $P_2 \rightarrow P_{45}$, and secondary link $S_1 \rightarrow S_{23}$ are active in time slot 2. Since all inter-network interference is canceled by the secondary nodes, underlay coexistence for the secondary nodes holds in time slot 2.

After our online distributed algorithm admits the new secondary session into the network, the

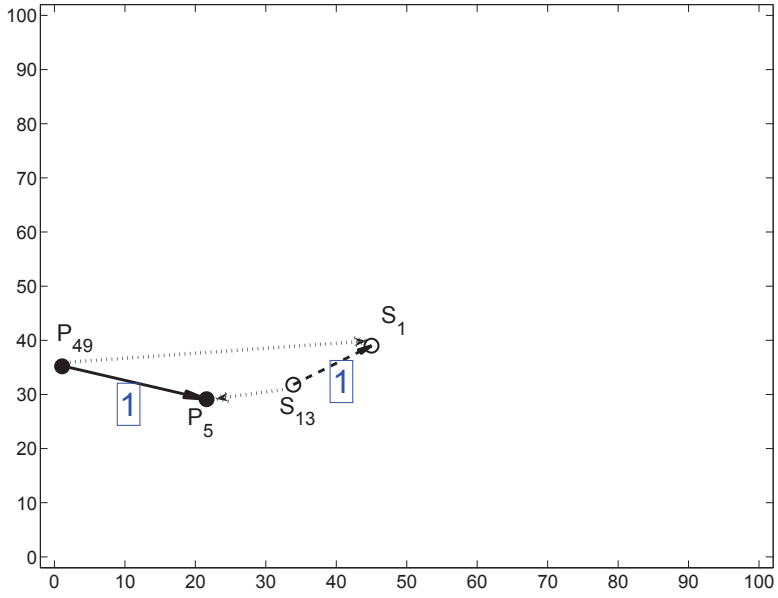


(a) Before the new secondary session arrives.

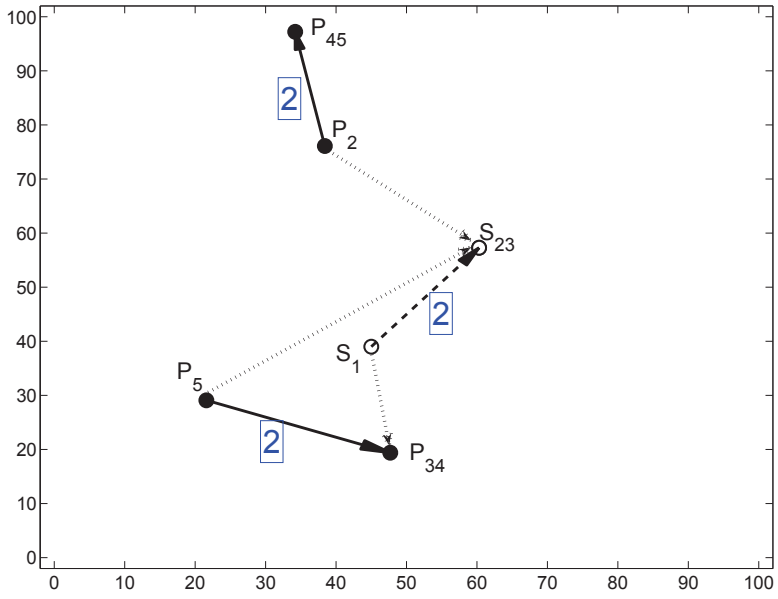


(b) After the new secondary session arrives.

Figure 4.10: The scheduling and routing before and after the new secondary session $S_{42} \rightarrow S_{10}$ arrives.

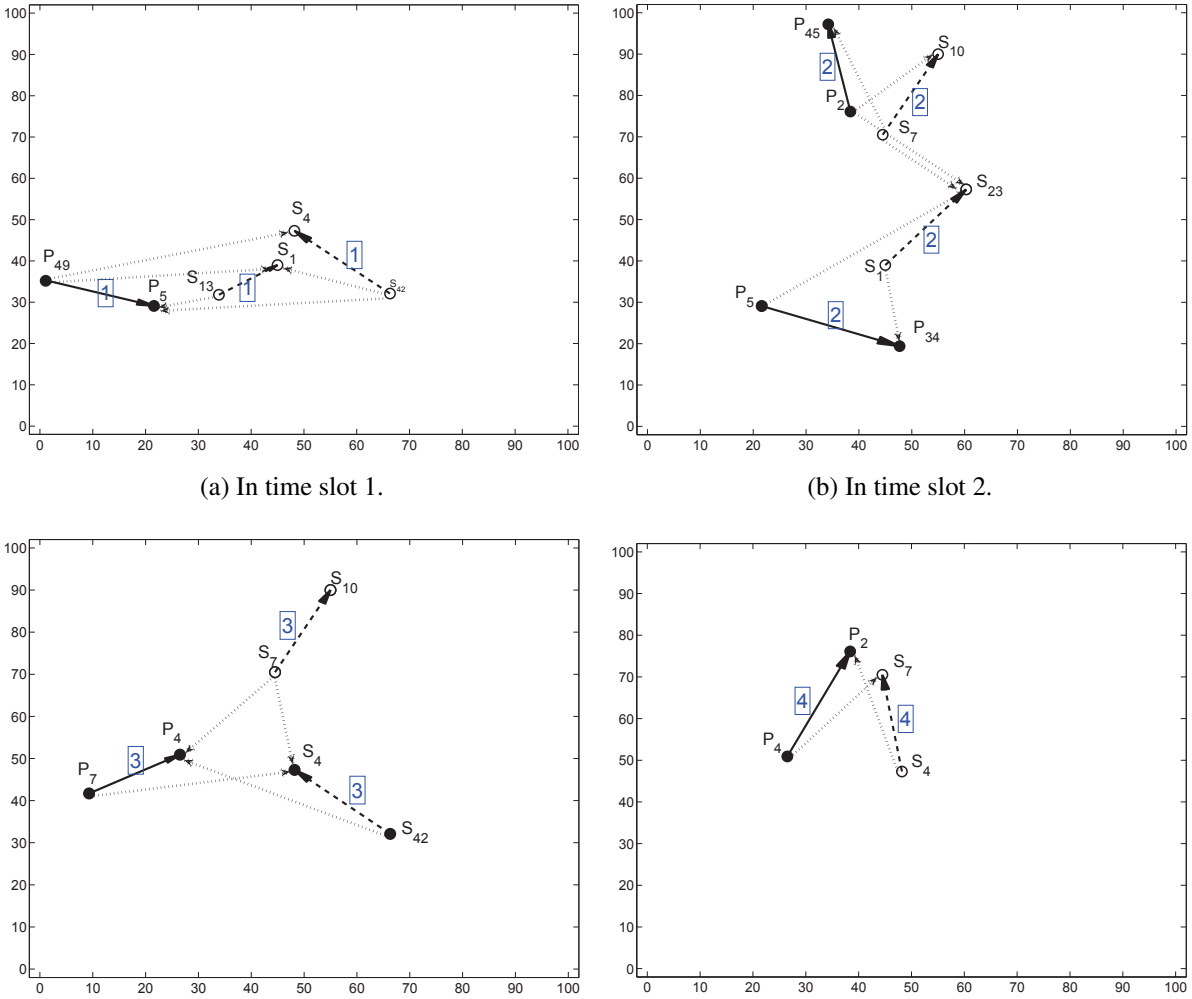



(a) In time slot 1.




(b) In time slot 2.

Figure 4.11: Interference relationship in the first two time slots before new secondary session $S_{42} \rightarrow S_{10}$ arrives.




(a) In time slot 1.

(b) In time slot 2.

(c) In time slot 3.

(d) In time slot 4.

Figure 4.12: Interference relationship in each time slot after new secondary session $S_{42} \rightarrow S_{10}$ arrives.

Table 4.2: DoF allocation for SM and IC for the secondary sessions in each time slot after the new session arrives.

| Time Slot 1 | | | | |
|---|---|---|---|---|
| Node $i$ | TX/RX | DoF for SM | DoF for IC to/from primary nodes | DoF for IC within secondary network |
| $S_{13}$ | TX | 2 | 1 to $P_5$ | NO |
| $S_1$ | RX | 2 | 1 from $P_{49}$ | NO |
| $S_{42}$ | TX | 1 | 1 to $P_5$ | 2 to $S_1$ |
| $S_4$ | RX | 1 | 0 | 2 from $S_{13}$ |
| Time Slot 2 | | | | |
| $S_1$ | TX | 2 | 1 to $P_{34}$ | NO |
| $S_{23}$ | RX | 2 | 1 from $P_2$, 1 from $P_5$ | NO |
| $S_7$ | TX | 1 | 1 to $P_{45}$ | 2 to $S_{23}$ |
| $S_{10}$ | RX | 1 | 1 from $P_2$ | No |
| Time Slot 3 | | | | |
| $S_{42}$ | TX | 1 | 1 to $P_4$ | NO |
| $S_4$ | RX | 1 | 1 from $P_7$ | NO |
| $S_7$ | TX | 1 | 1 to $P_4$ | 1 to $S_4$ |
| Time Slot 4 | | | | |
| $S_4$ | TX | 1 | 1 to $P_2$ | NO |
| $S_7$ | RX | 2 | 1 from $P_4$ | NO |

scheduling and routing for the primary and secondary sessions are shown in Fig. 4.10(b). Table 4.2 shows the details of DoF allocation for SM and IC at each secondary node, where the shaded rows correspond to those secondary nodes that are active before the arrival of this new secondary session. Comparing to Table 4.1, the DoF allocation for the shaded rows are not changed by our online algorithm. The interference relationships in each time slot is shown in Figs. 4.12(a), (b), (c) and (d). By cross-referencing the detailed information in Table 4.2, it is easy to verify, as we did for Fig. 4.11 and Table 4.1, that all inter-network and intra-network interferences are canceled by the secondary nodes. Therefore, the underlay coexistence holds in all time slots.

Following the same validation methodology, we have verified that underlay coexistence indeed

holds at all time instances (that we randomly pick for examination) for all possible arrival/departure events. Therefore, we conclude the our online algorithm can guarantee underlay coexistence.

## 4.7 Chapter Summary

The underlay paradigm allows extremely efficient utilization of spectrum by allowing simultaneous activation of the secondary nodes with the primary nodes. Such simultaneous activity is made possible through aggressive IC by the secondary nodes without any noticeable burden on the primary nodes. An effective online traffic management and IC algorithm is crucial for the secondary nodes to achieve underlay coexistence. In this chapter, we propose an online distributed algorithm to handle traffic dynamics for multi-hop primary and secondary networks. For IC, we employ MIMO at each secondary nodes and relied on the DoF allocation at each secondary node for IC. Through distributed computation and DoF resource allocation, we show that all inter-network and intra-network interference can be effectively canceled by the secondary nodes so that so that data transport is free of interference in both the primary and secondary networks. More important, we prove that such inter-network and intra-network IC through our DoF allocation is indeed feasible at the PHY layer at all time under traffic dynamics. By conducting performance evaluation under various traffic loads, we find that our online algorithm offers competitive performance when compared to an offline centralized algorithm.

# Chapter 5

# Policy-based Network Cooperation: Mathematical Modeling and Optimization

## 5.1 Introduction

The last decade has witnessed rapid advance in the research and development of spectrum-sharing technologies. Recent report by the President's Council of Advisors on Science and Technology (PCAST) [46] called for the sharing of 1 GHz of federal government radio spectrum with non-government entities in order to spur economic growth. This report further accelerated the pace of commercialization of innovative spectrum-sharing technologies. In [22], Goldsmith *et al.* outlined three spectrum-sharing paradigms for cognitive radios (CR), namely *underlay, overlay,* and *interweave.* These three paradigms were defined from an *information theoretic* perspective, solely based on how much side information (e.g., channel conditions, codebooks) is available to the CRs. In the networking community, these three paradigms have been mapped into specific scenarios of how primary and secondary networks interact with each other for data forwarding. Specifically, the *interweave* paradigm refers to the simple idea that secondary users are allowed to use a spectrum band allocated to the primary users only when the primary users are not using the band [3, 4, 14, 21, 26, 72, 89]. This paradigm is analogous to the classic interference avoidance

in medium access, or in CR terminology, dynamic spectrum access (DSA). This is the prevailing scenario on which most of research efforts have been devoted by the CR community in recent years.

The *underlay* paradigm refers to that secondary users' activities or interference on primary users is negligible (or below a given threshold). In contrast to the interweave paradigm, secondary users may be active *concurrently* with the primary users in the same vicinity and in the same frequency. Potential interference from the secondary users may be properly canceled (by the secondary users) via various interference cancelation (IC) techniques so that residual interference are negligible to the primary users [23, 33, 76, 85, 86].

Finally, the *overlay* paradigm requires that the secondary users have the primary users' codebook and messages so that the secondary users can help maintain or improve the communication of the primary users while still achieving some communication on their own. This is accomplished through sophisticated signal processing and coding (e.g., dirty paper coding (DPC) [15, 71] and power allocation [37]). From a networking perspective, the overlay paradigm can be interpreted as having secondary users help forward traffic of the primary users on top of its own communications.

Under the interweave and underlay paradigms, the primary and secondary networks are independent (in terms of data forwarding in each network). On the other hand, under the overlay paradigm, there is some level of cooperation by the secondary network. Inspired by this primitive cooperation idea in the overlay paradigm, there have been some recent efforts [28, 31, 42, 43, 56, 61, 83] on how to exploit possible cooperation from secondary users for the benefit of data forwarding. We will review these efforts in detail in Section 5.2. To summarize, the focus of these efforts has been limited to having secondary nodes help relay primary nodes' traffic. This, as we envision in this chapter, is only a tip of the iceberg.

In this chapter, we develop a paradigm with a much broader vision beyond the state of the art. We explore *network cooperation* as a new dimension for spectrum sharing between primary and secondary nodes. Such network cooperation can be defined as a set of *policies* under which different degrees of cooperation are to be achieved. Corresponding to each cooperation policy, a

traffic-forwarding behavior for primary and secondary users can be defined. One such primitive policy, as that in [28, 31, 42, 43, 56, 61, 83], is to have secondary network help relay primary users' traffic. Another policy (UPS [79]), which we will use as a main policy example in this chapter, is to allow complete node-level cooperation between the primary and secondary networks for data forwarding. These two examples are among many possible policies that one can define to achieve network sharing between primary and secondary networks.

To concretize our discussion on policy-based network sharing, we consider the UPS policy in detail, where UPS is the abbreviation of United cooperation of Primary and Secondary networks [79]. UPS represents a policy that allows a complete cooperation between the primary and secondary networks to relay each other's traffic. For performance evaluation, we study a problem with the goal of supporting the rate requirements of the primary sessions while maximizing the throughput of the secondary sessions. A number of technical challenges must be addressed in this problem, including how to provide guaranteed service for the primary traffic while supporting as much the secondary traffic as possible, how to select the optimal relays and routing paths for each source and destination pair, and how to coordinate the transmission and interference relationship between the primary and secondary nodes. For this problem, we develop an optimization model and formulate a combinatorial optimization problem. Although the problem is in the form of mixed-integer nonlinear program (MINLP), we develop an approximation solution based on the piece-wise linearization technique that allows to transform this problem into a mixed-integer linear program (MILP). Through simulation results, we demonstrate that UPS policy offers significantly better throughput performance than that under the existing interweave paradigm.

The remainder of this chapter is organized as follows. In Section 5.2, we review related work on primary and secondary network cooperation. In Section 5.3, we outline our vision of policy-based network cooperation and use UPS as an example. In Section 5.4, we use UPS as a case study for performance evaluation. For UPS, we develop an optimization model and formulate an optimization problem. In Section 5.5, we propose an approximation solution for the UPS throughput optimization problem. Section 6.4 presents simulation results to demonstrate the benefits and advantages of the UPS policy. Section 6.5 concludes this chapter and points out future research

directions.

## 5.2 Related Work

Due to space limitation, we will focus our attention on recent research efforts related to primary and secondary network cooperation. We find that all these efforts only considered having the secondary network help relay traffic for the primary network. In [61], Simeone *et al.* proposed to have the primary network lease its spectrum in the time domain to the secondary network in exchange for having the secondary network relay its data. In [83], Zhang and Zhang formulated this model as a Stackelberg game and a unique Nash Equilibrium point was achieved for maximizing primary and secondary users' utilities in terms of their transmission rates and revenue. In [56], Su *et al.* proposed to have the primary network lease its spectrum in the frequency domain to the secondary network to relay its data in order to maximize primary users' energy saving and secondary users' data rates. In [31], Jayaweera *et al.* proposed a new way to encourage primary users to lease their spectrum by having secondary users place bids on the amount of power they are willing to expend for relaying primary users' traffic. In [28], Hua *et al.* proposed a MIMO-based cooperative CR network where the secondary users utilize MIMO's antenna diversity to help relay primary users' traffic while transmitting their own traffic. In [42], Manna *et al.* considered the three-node model in [34]. The relay node was assumed to be a secondary node and have MIMO capability. The primary transmitter leases the second time slot to the secondary node (relay node) so that the secondary node can use the time slot to help relay the primary node's traffic while transmitting its own data. In [43], Nadkar *et al.* considered how to offer incentive (in terms of time and frequency) to a secondary network to help transmit primary user traffic. They studied a cross-layer optimization problem that maximizes transmission opportunities for secondary users while offering a guaranteed throughput to the primary users.

In all these efforts involving node-level cooperation between the primary and secondary networks, the focus has been limited to having secondary nodes help primary nodes in relaying pri-

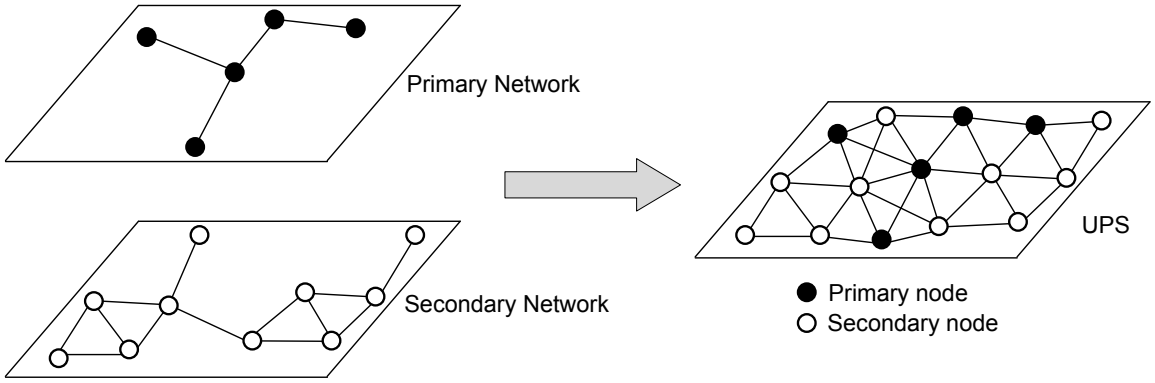



Figure 5.1: Network topologies under the interweave and the UPS policy.

mary users' traffic. As discussed, this is only a tip of the iceberg on network cooperation. In this chapter, we envision much broader cooperation between the two networks.

## 5.3 The Case of Policy-based Network Cooperation

As discussed in Section 7.1, the goal of this chapter is to outline a broad vision of policy-based network cooperation between the primary and secondary networks as a new dimension in radio spectrum sharing. Here, a policy defines the scope of cooperation at the node-level between the two networks. Such cooperation policies could vary from unilateral cooperation (i.e., only secondary nodes help relay primary user traffic but not vice versa), bilateral cooperation, constrained cooperation, or other customized policy based on particular application needs or requirements.

As a concrete example, we consider the UPS policy discussed in Section 7.1, which represents an interesting and extreme scenario where there is complete cooperation between the primary and secondary networks. Figure 5.1 illustrates the UPS policy for multi-hop primary and secondary networks. Unlike overlay, which is limited to only allowing secondary nodes help relay primary nodes' traffic, UPS allows primary nodes to help relay secondary nodes' traffic as well. From a network resource perspective, the UPS policy allows the pooling of all the resources from primary and secondary networks together and allows users in each network to access much richer network

resources in a combined network. Note that although the two networks are combined into one at the physical level, priority or service guarantee to the primary network traffic can still be enforced by implementing appropriate traffic engineering rules.

It is not hard to see that there are many potential benefits associated with the UPS policy. We briefly describe these benefits as follows:

- **Topology.** Comparing to having primary and secondary nodes being independent for each other, the combined network allows both primary and secondary networks a much improved connectivity with nodes from both networks.

- **Power Control.** As more nodes fall in the maximum transmission range of a primary or secondary node, this node has more flexibility in choosing its next hop node via power control. This flexibility can be exploited for different upper layer performance requirements or objectives.

- **Link Layer.** The improved physical topology allows more opportunities at the link layer for spectrum access. Both the primary and secondary networks can better coordinate with each other in transmission and interference avoidance. Further, the potential issue associated with link failure can now be mitigated effectively.

- **Network Diversity.** The combined network offers more routing opportunities to users in both networks. This directly translates into improved throughput and delay performance for user sessions.

- **Service and Applications.** The UPS architecture (combining both primary and secondary networks) allows to offer much richer services and applications than those services that were studied in [28, 31, 42, 43, 56, 61, 83]. Although the two networks are combined, the services and applications offered to users in each network can still be supported, by implementing certain traffic engineering policies. In other words, the combined network does not mean that service guarantee to the primary network will be lost. On the contrary, by specifying the desired resource management policy appropriately in the combined network, one can easily

achieve various service differentiation objectives and application goals, as we shall describe in a case study in the rest of this chapter.

The above UPS only represents one policy under the policy-based network cooperation paradigm. There are many other policies that can also be considered, ranging from no cooperation, unilateral cooperation, constrained cooperation, among others. Interweave and UPS can be considered two extreme cases of the policy space for network cooperation. The overlay paradigm that we discussed earlier (i.e., only secondary nodes helping primary traffic but not vice versa) may resemble the unilateral sharing policy, which can be viewed as a policy between the interweave and UPS. The *constrained cooperation* policy allows each network to only engage a subset of its nodes in network cooperation. The motivation of this policy is that certain nodes in either network may be too critical or sensitive (e.g., due to security concerns) in its own network and are thus prohibited from interacting with nodes from the other network. This constrained cooperation may be viewed as a generalization of interweave and UPS. Again, the policies discussed above only represent a few among a lot of possibilities. The definition of a policy is up to the network operators and it determines the scope of cooperation between the two networks.

The policy-based node-level cooperation paradigm may offer many possibilities and potential benefits for both the primary and secondary networks. From a networking perspective, the improved network connectivity, increased flexibility in power control, scheduling and routing all translate into improved forwarding performance for primary and secondary users' traffic. From a spectrum-sharing perspective, the ability to access other network infrastructure helps improve spatial diversity, thus allowing users to tap unused spectrum in the spatial domain. From economic perspective, such shared network infrastructure reduces the cost of infrastructure needed for each individual network (by allowing the tapping of another network's infrastructure resource), thus helping to enable traditionally underserved population and areas to benefit from current and future wireless-enabled goods and services. But from regulatory perspective, the proposed policy-based node-level cooperation paradigm may be ahead of its time. But there is no reason why we should not investigate its capability and recognize its potential from a research perspective. This is the goal of this chapter.

## 5.4 Case Study: UPS Policy

### 5.4.1 Problem Statement

In the rest of this chapter, we offer an in-depth study of the UPS policy. Referring to Fig. 5.1, suppose that there is a set of sessions in the primary network, with each session having a certain rate requirement. In the secondary network, suppose there is also a set of sessions, with each session having an elastic traffic requirement. By "elastic", we mean that each secondary session does not have a stringent rate requirement as the primary session. Instead, each secondary session will be supported on a best-effort basis and will transmit as much as the remaining network resource allows. A plausible goal under the UPS policy could be to have the combined network to support the rate requirements of the primary sessions while maximizing the throughput of the secondary sessions.

For this problem, there are a number of technical challenges that one must address:

- **Guaranteed service for primary traffic.** Since each primary session is assumed to have a hard rate requirement, the combined network should support it at all possibility. This problem alone may not be challenging. What is challenging (and interesting) is that should there are multiple ways to support primary sessions' rate requirements. We should find such a way that the rates for the secondary sessions are maximized in the combined network.

- **Relay selection.** To meet the service requirement (guaranteed service for primary traffic) and to optimize the objective (maximize the rates of secondary sessions), relay node selection along a route (for either a primary or secondary session) is not a trivial problem.

- **Scheduling.** To maximize the rates of the secondary sessions while guaranteeing the rates of the primary sessions, scheduling in each time slot needs to be carefully designed. In particular, in addition to addressing traditional self-interference (half-duplex) and mutual-interference problems, the primary network must be cooperative so as to help the secondary sessions to achieve their optimization objective in the combined network. Such cooperative

behavior from the primary network is a key in the UPS policy and has not been explored in prior efforts.

### 5.4.2 Mathematical Modeling

In this section, we develop a mathematical model for the UPS policy. Table 7.1 lists notation in this chapter. Denote $\mathcal{N}$ as the combined set of nodes consisting the set of primary nodes $\hat{\mathcal{N}}_{\rm P}$ and the set of secondary nodes $\mathcal{N}_{\rm S}$, i.e., $\mathcal{N} = \hat{\mathcal{N}}_{\rm P} \bigcup \mathcal{N}_{\rm S}$. In the combined network, denote $\mathcal{T}_i$ as the set of nodes (including both primary and secondary nodes) that is located within a nodes $i$'s transmission range, where $i$ can be either a primary or secondary node (i.e., $i \in \mathcal{N}$). Denote $\mathcal{I}_j$ as the set of nodes (including both primary and secondary nodes) that is located within node $j$'s interference range, where $j$ can be either a primary or secondary node. For a primary session $l \in \hat{\mathcal{L}}$, we assume it has a hard requirement on its data rate, which we denote as $\hat{R}(l)$. For a secondary session $m \in \mathcal{L}$, we assume that it does not have a rate requirement. Instead, the data rate $r(m)$ on $m \in \mathcal{L}$ is supported on a best-effort basis and will be an optimization variable in the problem formulation.

**Guaranteed Service for the Primary Sessions.** For primary sessions, they consider the combined network $\mathcal{N}$ as their communication resources. For flexibility and load balancing, we allow flow splitting in the network. That is, the flow rate of a session may split and merge inside the network in whatever loop-free manner as long as it can help support the given rate requirement $\hat{R}(l)$ of session $l \in \hat{\mathcal{L}}$. Denote $\hat{f}_{ij}(l)$ as the data rate on link $(i, j)$ that is attributed to primary session $l \in \hat{\mathcal{L}}$, where $i \in \mathcal{N}$ and $j \in \mathcal{T}_i$. Denote $\hat{s}(l)$ and $\hat{d}(l)$ as the source and destination nodes of primary session $l \in \hat{\mathcal{L}}$, respectively. We have the following flow balance constraints:

- If node $i$ is the source node of primary session $l \in \hat{\mathcal{L}}$ (i.e., $i = \hat{s}(l)$), then

$$\sum_{j \in \mathcal{T}_i} \hat{f}_{ij}(l) = \hat{R}(l) \qquad (l \in \hat{\mathcal{L}}). \tag{5.4.1}$$

Table 5.1: Notation for UPS paradigm

| | |
|---|---|
| Primary Network | |
| $\hat{\mathcal{N}}_{\mathrm{P}}$ | The set of primary nodes |
| $\hat{\mathcal{L}}$ | The set of primary sessions |
| $\hat{f}_{ij}(l)$ | The flow rate traversing on link $(i,j)$ that is attributed to primary session $l \in \hat{\mathcal{L}}, i, j \in \mathcal{N}$ |
| $\hat{s}(l)$ | The source node of primary session $l \in \hat{\mathcal{L}}$ |
| $\hat{d}(l)$ | The destination node of primary session $l \in \hat{\mathcal{L}}$ |
| $\hat{R}(l)$ | The data rate requirement of primary session $l \in \hat{\mathcal{L}}$ |
| Secondary Network | |
| $\mathcal{N}_{\mathrm{S}}$ | The set of secondary nodes |
| $\mathcal{L}$ | The set of secondary sessions |
| $f_{ij}(m)$ | The flow rate traversing on link $(i,j)$ that is attributed to secondary session $m \in \mathcal{L}, i, j \in \mathcal{N}$ |
| $s(m)$ | The source node of secondary session $m \in \mathcal{L}$ |
| $d(m)$ | The destination node of secondary session $m \in \mathcal{L}$ |
| $r(m)$ | The data rate achieved by secondary session $m \in \mathcal{L}$ |
| Combined Network | |
| $\mathcal{N}$ | The set of all nodes in the network, $\mathcal{N} = \hat{\mathcal{N}}_{\mathrm{P}} \bigcup \mathcal{N}_{\mathrm{S}}$ |
| $C_{ij}$ | The link capacity of link $(i,j), i, j \in \mathcal{N}$ |
| $x_{ij}[t]$ | $= 1$ if node $i$ is transmitting data to node $j$ in time slot $t$, and is $0$ otherwise |
| $\mathcal{T}_i$ | The set of nodes that are located within the transmission range of node $i \in \mathcal{N}$ |
| $\mathcal{I}_i$ | The set of nodes that are located within the interference range of node $i \in \mathcal{N}$ |
| $T$ | The number of time slots in a frame |

- If node $i$ is an intermediate relay node for primary session $l$ (i.e., $i \neq \hat{s}(l)$ and $i \neq \hat{d}(l)$), then

$$\sum_{j \in \mathcal{T}_i}^{j \neq \hat{s}(l)} \hat{f}_{ij}(l) = \sum_{k \in \mathcal{T}_i}^{k \neq \hat{d}(l)} \hat{f}_{ki}(l) \qquad (l \in \hat{\mathcal{L}}, i \in \hat{\mathcal{N}}_{\mathrm{P}}). \tag{5.4.2}$$

- If node $i$ is the destination node of primary session $l$ (i.e., $i = \hat{d}(l)$), then

$$\sum_{k \in \mathcal{T}_i} \hat{f}_{ki}(l) = \hat{R}(l) \qquad (l \in \hat{\mathcal{L}}). \tag{5.4.3}$$

It can be easily verified that once (5.4.1) and (5.4.2) are satisfied, then (5.4.3) is also satisfied. As a result, it is sufficient to list only (5.4.1) and (5.4.2) in the formulation.

**Best-effort Service for Secondary Sessions.** Under the UPS policy, the primary sessions have priority in access the combined network resources (in the form of guaranteed services). Once the primary sessions are supported, the secondary sessions may use as much as the remaining resources in the combined network. How the primary and secondary sessions interact in the combined network should be part of an optimization problem. Denote $f_{ij}(m)$ as the data rate on link $(i, j)$ that is attributed to secondary session $m \in \mathcal{L}$. Denote $s(m)$ and $d(m)$ as the source and destination nodes of secondary session $m \in \mathcal{L}$, respectively. Similar to that for the primary sessions, we allow flow splitting for the secondary sessions. We have the following flow balance constraints:

- If node $i$ is the source node of secondary session $m \in \mathcal{L}$ (i.e., $i = s(m)$), then we have

$$\sum_{j \in \mathcal{T}_i} f_{ij}(m) = r(m) \qquad (m \in \mathcal{L}), \tag{5.4.4}$$

- If node $i$ is an intermediate relay node for secondary session $m$ (i.e., $i \neq s(m)$ and $i \neq d(m)$), then

$$\sum_{j \in \mathcal{T}_i}^{j \neq s(m)} f_{ij}(m) = \sum_{k \in \mathcal{T}_i}^{k \neq d(m)} f_{ki}(m) \qquad (m \in \mathcal{L}, i \in \mathcal{N}_{\mathrm{S}}), \tag{5.4.5}$$

- If node $i$ is the destination node of secondary session $m$ (i.e., $i = d(m)$), then

$$\sum_{k \in \mathcal{T}_i} f_{ki}(m) = r(m) \qquad (m \in \mathcal{L}). \tag{5.4.6}$$

Again, to avoid redundancy, it is sufficient to list only (5.4.4) and (5.4.5) in the formulation.

Note that although (5.4.4)–(5.4.6) are similar to (5.4.1)–(5.4.3), there is an important difference between them: unlike $\hat{R}(l)$ for primary session $l \in \hat{\mathcal{L}}$, which is a given *constant*, secondary session rate $r(m)$, $m \in \mathcal{L}$, is an optimization *variable*. Therefore, for the primary sessions, we only need to optimize their flow paths, while for the secondary sessions, we need to optimize both their routes and their rates.

**Self-interference Constraints.** We assume scheduling is done in time slot on a frame-by-frame basis, with each frame consisting of $T$ time slots. We use a binary variable $x_{ij}[t], i, j \in \mathcal{N}$ and $1 \leq t \leq T$, to indicate whether node $i$ transmits data to node $j$. That is,

$$x_{ij}[t] = \begin{cases} 1 & \text{If node } i \text{ transmits data to node } j \\ & \text{in time slot } t; \\ 0 & \text{otherwise.} \end{cases}$$

where $i \in \mathcal{N}, j \in \mathcal{T}_i$, and $1 \leq t \leq T$.

Assuming each primary or secondary session is unicast, a node $i$ only needs to transmit to or receive from one node in a time slot. We have

$$\sum_{j \in \mathcal{T}_i} x_{ij}[t] \leq 1 \qquad (i \in \mathcal{N}, 1 \leq t \leq T) \ , \tag{5.4.7}$$

$$\sum_{k \in \mathcal{T}_i} x_{ki}[t] \leq 1 \qquad (i \in \mathcal{N}, 1 \leq t \leq T) \ . \tag{5.4.8}$$

To account for half-duplex at each node $i$, we have:

$$x_{ij}[t] + x_{ki}[t] \leq 1 \qquad (i \in \mathcal{N}, j, k \in \mathcal{T}_i, 1 \leq t \leq T) \ . \tag{5.4.9}$$

These three constraints in (5.4.7), (5.4.8) and (5.4.9) can be replaced by the following single constraint.

$$\sum_{j \in \mathcal{T}_i} x_{ij}[t] + \sum_{k \in \mathcal{T}_i} x_{ki}[t] \leq 1 \qquad (i \in \mathcal{N}, 1 \leq t \leq T). \tag{5.4.10}$$

To see this, note that in (5.4.10), if node $i$ is receiving data from some node in $\mathcal{T}_i$ in time slot $t$, we must have $\sum_{j\in\mathcal{T}_i} x_{ij}[t] = 0$, i.e., node $i$ cannot transmit in the same time slot. This is exactly the half-duplex constraint. In this case, (5.4.10) also becomes (5.4.8). On the other hand, if node $i$ is transmitting to some node in $\mathcal{T}_i$ in time slot $t$, then $\sum_{k\in\mathcal{T}_i} x_{ki}[t] = 0$, i.e., node $i$ cannot receive in the same time slot. Again, this is the half-duplex constraint. In this case, (5.4.10) becomes (5.4.7).

**Mutual Interference Constraints.** To model mutual interference constraints, we assume that for any primary or secondary node $j \in \mathcal{N}$ that is receiving data in time slot $t$, it shall not be interfered by another (unintended) transmitting node $p \in \mathcal{I}_j$ in the same time slot. We have the following mutual interference constraint:

$$x_{ij}[t] + x_{pk}[t] \leq 1 \ , \tag{5.4.11}$$

where $i \in \mathcal{T}_j, p \in \mathcal{I}_j, k \in \mathcal{T}_p, j \in \mathcal{N}, j \neq k$, and $1 \leq t \leq T$.

Following the same token in (5.4.10), the three constraints in (5.4.7), (5.4.8) and (5.4.11) can be replaced by the following single and equivalent constraint.

$$\sum_{i\in\mathcal{T}_j} x_{ij}[t] + \sum_{k\in\mathcal{T}_p} x_{pk}[t] \leq 1 \ , \tag{5.4.12}$$

where $p \in \mathcal{I}_j, j \in \mathcal{N}, j \neq k$, and $1 \leq t \leq T$.

**Link Rate Constraints.** For each link $(i, j)$, denote the link capacity as $C_{ij}$, e.g., $C_{ij} = B\log_2(1 + \frac{Q_i d_{ij}^{-\alpha}\lambda}{N_0})$, where $B$ is bandwidth, $Q_i$ is the power spectral density from transmit node $i$, $d_{ij}$ is the distance between node $i$ and $j$, $\alpha$ is the path loss index, $\lambda$ is the antenna related constant, and $N_0$ is the ambient Gaussian power spectral density. Since the aggregate flow rate from the primary and secondary sessions on each link $(i, j)$ cannot exceed the average link rate (over $T$ time slots), we have

$$\sum_{l\in\hat{\mathcal{L}}}^{j\neq\hat{s}(l),i\neq\hat{d}(l)} \hat{f}_{ij}(l) + \sum_{m\in\mathcal{L}}^{j\neq s(m),i\neq d(m)} f_{ij}(m) \leq \frac{1}{T}\sum_{t=1}^{T} C_{ij}\cdot x_{ij}[t]. \tag{5.4.13}$$

### 5.4.3 Problem Formulation

In the combined network, our goal is to offer guaranteed support for the primary sessions (each with a given rate requirement) while maximizing the throughput for the secondary sessions, whose traffic is assumed to be elastic. For maximizing secondary network throughput, different objective functions can be explored to satisfy network requirement. In [79], we considered a simple case with linear objective function (i.e., maximizing the minimum throughput). In this chapter, we will consider a nonlinear objective function. We use a utility function $\ln r(m)$ for $m \in \mathcal{L}$ as our objective. Such utility function is widely used in the literature [48]. We have the following problem formulation:

OPT

max $\sum_{m \in \mathcal{L}} \ln r(m)$

s.t. Guaranteed service for primary sessions: (5.4.1), (5.4.2);
Best effort service for secondary sessions: (5.4.4), (5.4.5);
Self interference constraints: (5.4.10);
Mutual interference constraints: (5.4.12);
Link capacity constraints: (5.4.13);

In this formulation, $\hat{R}(l)$ and $C_{ij}$ are constants, $x_{ij}[t]$ are binary variables, $\hat{f}_{ij}(l)$, $f_{ij}(m)$ and $r(m)$ are continuous variables. Due to nonlinear terms $\ln r(m)$ in the objective function and binary variables $x_{ij}[t]$, the optimization problem is a *mixed-integer nonlinear programming* (MINLP), which is NP-hard in general. In the next section, we develop an approximation algorithm to solve this problem.

## 5.5 An Approximate Solution

### 5.5.1 Overview

In this section, we develop an approximate solution to OPT with guaranteed performance. For the nonlinear *log* term in the objective function of OPT, one could relax the nonlinear function with a series of linear functions. The issue here is how to achieve such linearization with performance guarantee. This is the focus of our proposed solution.

For a target performance gap $\epsilon$ between the optimal objective (unknown) and the approximate objective (that we aim to develop), we will develop an algorithm to determine a set of piece-wise linear segments that approximate the log function (See Fig. 5.2). The essence of our linear approximation is to find just the right number of linear and unequal-length segments to approximate the log function. The idea is that, for a given performance gap $\epsilon$, we can calculate the maximum linear approximation error, say $\eta$, that is allowed in the linearization. Then, we can develop an algorithm (Section 5.5.2) to find the slopes and starting points for the set of linear segments. Subsequently, the nonlinear log terms in OPT can be replaced by a set of linear constraints and we have a new linearized optimization problem, which we denote as OPT-L. Although OPT-L is in the form of *mixed-integer linear programming* (MILP), the integer variables are all binary. We find that commercial software (such as CPLEX) can solve such binary MILP efficiently.

### 5.5.2 Linearization

Our goal of linear approximation of $\ln r(m)$ is to replace $\ln r(m)$ with the minimum number of linear segments while ensuring that the difference between any point on $\ln r(m)$ and its corresponding linear segment is no more than $\eta$. Denote $K_m$ as the minimum number of line segments such that each segment meets the error requirement (i.e., $\eta$). Denote $r_L(m)$ and $r_U(m)$ as the lower and upper bounds for $r(m)$, respectively. For $r_L(m)$, we can set it to an arbitrarily small positive value. For $r_U(m)$, we can set it to $\max_{i,j\in\mathcal{N}} C_{ij}$, the maximum capacity among all links. Denote

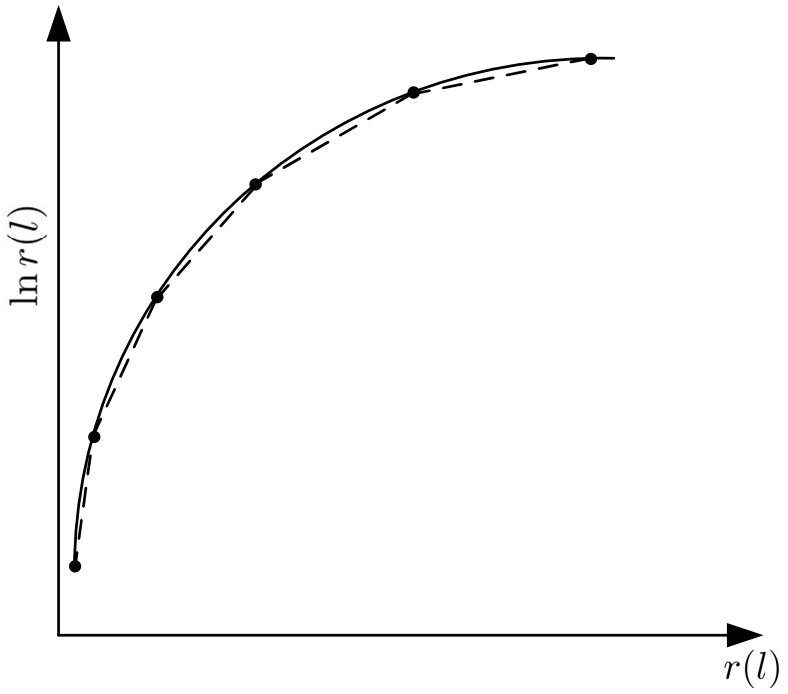



Figure 5.2: Piece-wise approximation with line segments.

$r_0(m), r_1(m), \cdots, r_{K_m}(m)$ as values on the X-axis for the end points of these $K_m$ segments, with $r_0(m) = r_L(m)$ and $r_{K_m}(m) = r_U(m)$.

The minimum number of line segments $K_m$ can be found with the following iterative process. We start from $r_0(m)$ to calculate the slope of the first segment, which must ensure that this segment satisfies the error bound $\eta$. After finding this slope, we can find the right-side end point of the first segment. From this point, we repeat the same process for the second segment and so forth, until the last segment exceeds $r_U(m)$.

Specifically, denote slope of the $k$-th linear segment as $q_k(m)$, i.e.,

$$q_k(m) = \frac{\ln r_k(m) - \ln r_{k-1}(m)}{r_k(m) - r_{k-1}(m)}. \tag{5.5.1}$$

Denote $y_k(r(m))$ as the $k$-th linear segment that approximates $\ln r(m)$. Then we have:

$$y_k(r(m)) = q_k(m) \cdot [r(m) - r_{k-1}(m)] + \ln r_{k-1}(m), \tag{5.5.2}$$

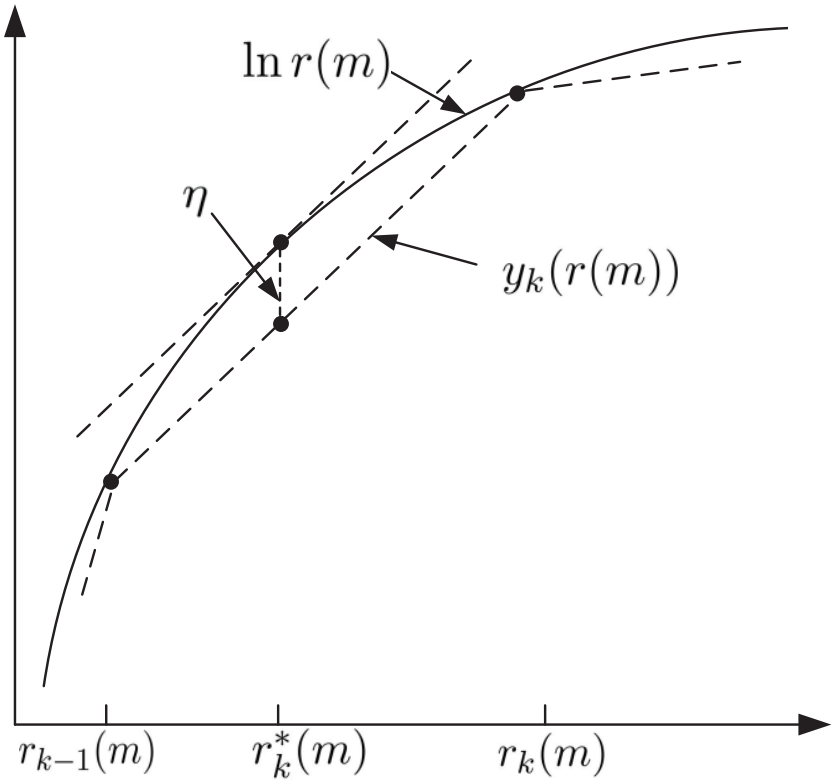



Figure 5.3: An illustration of the maximum approximation error for piece-wise line segment.

for $r_{k-1}(m) \leq r(m) \leq r_k(m)$.

Referring to Fig. 5.3, for any point $r(m)$ within $r_{k-1}(m) \leq r(m) \leq r_k(m)$, it is easy to see that the point on the tangential line (in parallel to the linear segment approximation) that intersects the $\log$ curve has the maximum approximation error $\eta$. Denote the X-coordinate of this point as $r_k^*(m)$, we have $\eta = \ln r_k^*(m) - y_k(r_k^*(m))$. Since the slope of tangential line (achieving $\eta$) for $\ln r(m)$ is $\frac{1}{r(m)}$, then $q_k(m) = \frac{1}{r_k^*(m)}$, or $r_k^*(m) = \frac{1}{q_k(m)}$, where $q_k(m)$ is the slope of the linear segment $y_k(r(m))$. Then, we have

$$\begin{aligned}
\eta =& \ln r_k^*(m) - y_k(r_k^*(m)) \\
=& \ln r_k^*(m) - \Big[ q_k(m) \cdot \Big( r_k^*(m) - r_{k-1}(m) \Big) + \ln r_{k-1}(m) \Big] \\
=& \ln \frac{1}{q_k(m)} - q_k(m) \cdot \left( \frac{1}{q_k(m)} - r_{k-1}(m) \right) - \ln r_{k-1}(m) \\
=& - \ln q_k(m) - 1 + q_k(m) \cdot r_{k-1}(m) - \ln r_{k-1}(m).
\end{aligned}$$

Therefore, we have the following equation:

$$-\ln q_k(m) + q_k(m) \cdot r_{k-1}(m) - [\ln r_{k-1}(m) + \eta + 1] = 0. \tag{5.5.3}$$

For a give error bound $\eta$, the values of $r_1(m), \cdots, r_{K_m}(m)$ and slopes $q_1(m), q_2(m), \cdots, q_{K_m}(m)$ can be found iteratively through the following algorithm:

**Algorithm 5.1. (Piece-wise Linearization)**

*Initialization:* $k := 0$ *and* $r_k(m) := r_L(m)$.

*While (*$r_k(m) < r_U(m)$*)* {

$k := k + 1$.

*Find slope* $q_k(m)$ *by solving the equation (5.5.3).*

*With* $q_k(m)$*, compute* $r_k(m)$ *via (5.5.1).* }

$K_m := k, r_{K_m}(m) := r_U(m)$.

*Recalculate* $q_{K_m}(m)$ *with (5.5.1).*

The values of $q_k(m)$ in (5.5.3) and $r_k(m)$ in (5.5.1) can be solved by numerical methods such as bisection method or Newton's method [52].

**Lemma 5.1.** *The maximum approximation error within each linear segment as defined by Algorithm 5.1 is no more than* $\eta$.

*Proof.* The proof is based on the above construction. We omit its discussion here to conserve space. □

### 5.5.3 Approximation Gap

By using the piece-wise linearization algorithm (Algorithm 5.1), we can approximate the log term $\ln r(m)$ with a series of linear segments, each with an approximation error no more than $\eta$. For $r(m)$, denote $y(m)$ as the concatenation of the piece-wise linear segments constructed by Algorithm 5.1. Then the objective function $\max \sum_{m \in \mathcal{L}} \ln r(m)$ in OPT is replaced by the following

linear objective and a set of linear constraints (representing the convex hull below the linear segments):

$$\max \quad \sum_{m\in\mathcal{L}} y(m) \tag{5.5.4}$$

$$s.t \quad y(m) \le q_k(m)\cdot(r(m) - r_{k-1}(m)) + \ln r_{k-1}(m) \tag{5.5.5}$$

$$(k = 1, 2, \cdots, K_m, m \in \mathcal{L}).$$

The original OPT can be re-formulated into a new optimization problem, which we denote as OPT-L.

OPT-L

max $\sum_{m\in\mathcal{L}} y(m)$

s.t. Constraints (5.4.1), (5.4.2), (5.4.4), (5.4.5), (5.4.10), (5.4.12), (5.4.13), (5.5.5),

$r_L(m) \le r(m) \le r_U(m)$, $(m \in \mathcal{L})$

$x_{ij}[t] \in \{0, 1\}, f_{ij}(m) \ge 0, \hat{f}_{ij}(l) \ge 0.$

$(i \in \mathcal{N}, j \in \mathcal{T}_i, m \in \mathcal{L}, l \in \hat{\mathcal{L}}, 1 \le t \le T)$,

where $x_{ij}[t]$ are binary variables, $\hat{f}_{ij}(l), f_{ij}(m), r(m)$ and $y(m)$ are continuous variables, and $q_k(m), r_{k-1}(m)$ and $\hat{R}(l)$ are constants. OPT-L is in the form of *mixed integer linear programming* (MILP). Since all integers are binary, the MILP problem tends to be solved efficiently by a commercial solver (CPLEX). Our simulation results in Section 6.4 confirm that this is indeed the case.

We now quantify the gap between the optimal objective values of OPT-L and OPT.

**Lemma 5.2.** *The gap between the optimal objective values of OPT and OPT-L, $\epsilon$, is upper bounded by $|\mathcal{L}| \cdot \eta$.*

*Proof.* Suppose an optimal solution of OPT is $\varphi^*_{\mathrm{OPT}} = [x^*_{ij}[t], r^*(m), f^*_{ij}(m), \hat{f}^*_{ij}(l)]$, with the objective value being $Y^*_{\mathrm{OPT}} = \sum_{m\in\mathcal{L}} \ln r^*(m)$. We can construct a feasible solution to OPT-L, denoted as $\varphi_{\mathrm{OPT-L}}$, based on $\varphi^*_{\mathrm{OPT}}$ as follows: $\varphi_{\mathrm{OPT-L}} = [x_{ij}[t], r(m), f_{ij}(m), \hat{f}_{ij}(l), y(m)]$, where $x_{ij}[t] = x^*_{ij}[t], r(m) = r^*(m), f_{ij}(m) = f^*_{ij}(m)$ and $\hat{f}_{ij}(l) = \hat{f}^*_{ij}(l)$. Then, $\varphi_{\mathrm{OPT-L}}$ satisfy

constraints (5.4.1), (5.4.2), (5.4.4), (5.4.5), (5.4.10), (5.4.12), (5.4.13) in OPT-L. $y(m)$ can be calculated by solving OPT-L with the variables being set to those values in $\varphi_{\mathrm{OPT-L}}$. Suppose that $r^*(m)$ falls in the interval $[r_{k-1}(m), r_k(m)]$. Then the objective function $\sum_{m\in\mathcal{L}} y(m)$ is maximized only when $y(m) = y_k(r^*(m)) = q_k(m) \cdot (r^*(m) - r_{k-1}(m)) + \ln r_{k-1}(m)$ in (5.5.5). Denote this objective value in OPT-L as $Y_{\mathrm{OPT-L}}$. Then,

$$\begin{aligned} Y^*_{\mathrm{OPT}} - Y_{\mathrm{OPT-L}} &= \sum_{m\in\mathcal{L}} \ln r^*(m) - \sum_{m\in\mathcal{L}} y(m) \\ &= \sum_{m\in\mathcal{L}} \ln r^*(m) - \sum_{m\in\mathcal{L}} y_k(r^*(m)) \\ &= \sum_{m\in\mathcal{L}} \Big[ \ln r^*(m) - y_k(r^*(m)) \Big] \\ &\leq |\mathcal{L}| \cdot \eta, \end{aligned}$$

where last inequality holds by Lemma 5.1. We let $\epsilon = |\mathcal{L}| \cdot \eta$.

Now denote $\varphi^*_{\mathrm{OPT-L}}$ as an optimal solution for OPT-L, with the objective value of $Y^*_{\mathrm{OPT-L}}$. Since $Y_{\mathrm{OPT-L}}$ is merely the objective value of a feasible solution, we have $Y^*_{\mathrm{OPT-L}} \geq Y_{\mathrm{OPT-L}}$. Then $Y^*_{\mathrm{OPT}} - Y^*_{\mathrm{OPT-L}} \leq Y^*_{\mathrm{OPT}} - Y_{\mathrm{OPT-L}} \leq \epsilon$. This completes the proof. □

Our complete solution for solving OPT can be summarized as follows: For any a given performance gap $\epsilon$, we can compute linear approximation error $\eta = \frac{\epsilon}{|\mathcal{L}|}$. Based on the approximation error $\eta$, we perform piece-wise linear approximation through Algorithm 5.1. Then we reformulate OPT to OPT-L, and solve it by CPLEX.

## 5.6 Simulation Results

In this section, we present numerical results to demonstrate the capabilities and advantages of the UPS policy. The goal of this section is twofold. First, we show that the UPS policy offers much better performance for both the primary and secondary networks than that under the interweave paradigm. Second, we shall have a close look at how the primary and secondary nodes help each other in the UPS policy.

### 5.6.1 Simulation Setting

We consider a UPS network where both the primary and the secondary nodes are randomly deployed in a $100 \times 100$ area. For generality, we normalize the units for distance, bandwidth, power and data rate with appropriate dimensions. We assume the bandwidth of the channel allocated to the primary network is $B = 10$. The number of time slots in a frame is $T = 10$. The transmission power spectral density $Q_i$ for each node $i \in \mathcal{N}$ is 1, the path loss index is $4$, the antenna related constant $\lambda$ is $1$, and the ambient Gaussian power spectral density $N_0 = 10^{-6}$. We assume the transmission range and interference range at all nodes are 30 and 50, respectively.

We set the maximum acceptable performance gap between the objective of OPT and its linear approximation OPT-L as $\epsilon = 0.02$.

### 5.6.2 An Example

We consider a 30-node network, with 15 primary nodes and 15 secondary nodes randomly deployed in a $100 \times 100$ area (see Fig. 5.4). The location of each node is given in Table 5.2. In this example, we assume that there are two primary sessions in the primary network and two secondary sessions in the secondary network. The source and destination nodes for each session are randomly chosen in each network and are shown in Table 5.3. Denote the rate requirements of the two primary sessions as $\hat{R}(1)$ and $\hat{R}(2)$, respectively. We gradually increase the rate requirements of $\hat{R}(1)$ and $\hat{R}(2)$ and examine (i) whether such rates can be supported under the UPS policy and the interweave paradigm, respectively, and (ii) the objective values of secondary session utilities under both the UPS policy and the interweave paradigm. The utility maximization problem for the secondary sessions under the interweave paradigm can be formulated following a similar token to OPT.

Table 5.4 shows the approximation gap between the utility objective of the linearized problem and the utility objective of the original problem under different $\hat{R}(1)$ and $\hat{R}(2)$. The first column represents increasing rate requirements for the primary sessions. The second column shows the

Table 5.2: Location of primary and secondary nodes for the 30-node network.

| Primary Node | Location | Secondary Node | Location |
|---|---|---|---|
| $P_1$ | (2.5, 85.2 ) | $S_1$ | (29.6, 76.6) |
| $P_2$ | (29.2, 95.5 ) | $S_2$ | (55.5, 62 ) |
| $P_3$ | (11.4, 59.1 ) | $S_3$ | (50.4, 97.1) |
| $P_4$ | (45.9, 79) | $S_4$ | (70.7, 62.2) |
| $P_5$ | (63.8, 67.8) | $S_5$ | (19.1, 87.4 ) |
| $P_6$ | (54, 41.2) | $S_6$ | (62, 38.4) |
| $P_7$ | (86.3, 56.5) | $S_7$ | (77, 26.2) |
| $P_8$ | (68.4, 87.5) | $S_8$ | (43.4, 40.8 ) |
| $P_9$ | (34, 56.3) | $S_9$ | (92.4, 44.1) |
| $P_{10}$ | (78.3, 41.7) | $S_{10}$ | (70.7, 6.6)) |
| $P_{11}$ | (33.5, 19.6 ) | $S_{11}$ | (20.1, 46.1) |
| $P_{12}$ | (79, 83.7) | $S_{12}$ | (92.3, 74.8 ) |
| $P_{13}$ | (95.9, 31.5) | $S_{13}$ | (88, 96.4) |
| $P_{14}$ | (19.5, 30.1) | $S_{14}$ | (2.4, 29) |
| $P_{15}$ | (54.4, 13.8) | $S_{15}$ | (92.6, 8.6) |

Table 5.3: The source and destination nodes for each session in the 30-node network.

| Session | Source | Destination |
|---|---|---|
| Primary session 1 | $P_{13}$ | $P_{11}$ |
| Primary session 2 | $P_3$ | $P_8$ |
| Secondary session 1 | $S_{13}$ | $S_6$ |
| Secondary session 2 | $S_{14}$ | $S_2$ |

utility objectives of the two secondary sessions (abbreviated as “SS” in the table) from the linearized problem OPT-L, while the third column shows the utility objectives of the two secondary

sessions from the original problem. The fourth column shows the gap between the utility objectives from the linearized problem and the original problem. Given the target approximation error $\varepsilon = 0.02$, all actual approximation errors fall below this target.

Table 5.5 summarizes the results of this study. The second column represents increasing rate requirements for the primary sessions (i.e., $\hat{R}(1) = \hat{R}(2)$). For ease of explanation, we break this table into five regions, with each region representing a specific behavior for comparison between the UPS policy and interweave paradigm. The third and fourth columns show the performance under the UPS policy. Specifically, the third column shows whether the rate requirements of the two primary sessions can be supported ("feasible") in the primary network (abbreviated as "PN" in the table); the fourth column shows the rate utility objective of the two secondary sessions (abbreviated as "SS" in the table) with $-\infty$ indicating zero rates for the secondary sessions (due to the log function) and "N/A" indicating not applicable as the corresponding network cannot even support the rate requirements of the primary sessions. The fifth and sixth columns show the performance under the interweave paradigm, which are to be compared to the third and fourth columns under the UPS policy, respectively.

**Region 1** This region represents the scenario where the rate requirements of the primary sessions can be supported under both the UPS policy and the interweave paradigm, and the rates of the secondary sessions are positive. Comparing columns four and six, we can find that the secondary sessions always achieve higher utility objectives under the UPS policy than that under the interweave paradigm. This confirms our expectation that the UPS policy can offer higher throughput for the secondary sessions.

As an example, consider the case when both the two primary sessions have rate requirements 1.6. The utility objectives achieved for the secondary sessions under the UPS policy and the interweave paradigms are 3.288 and 1.263, respectively. Specifically, the rates for the two secondary sessions are 4.784 and 5.692 under the UPS policy while the rates for the same two secondary sessions are 1.776 and 2.024 under the interweave paradigm. Under the UPS policy, the flow routing and scheduling for the primary and secondary sessions are shown in Fig. 5.4(a). The number in

Table 5.4: The approximation gap between the SS utility objectives of linearized problem and original problem.

| Rate Requirement $\hat{R}(1), \hat{R}(2)$ | SS Utility of Linearized Problem | SS Utility of Original Problem | Gap |
|---|---|---|---|
| 0 | 3.7012 | 3.7128 | 0.0016 |
| 0.2 | 3.288 | 3.3046 | 0.0016 |
| 0.4 | 3.288 | 3.3046 | 0.0016 |
| 0.6 | 3.288 | 3.3046 | 0.0016 |
| 0.8 | 3.288 | 3.3046 | 0.0016 |
| 1.0 | 3.288 | 3.3046 | 0.0016 |
| 1.2 | 3.288 | 3.3046 | 0.0016 |
| 1.4 | 3.288 | 3.3046 | 0.0016 |
| 1.6 | 3.288 | 3.3046 | 0.0016 |
| 1.8 | 3.158 | 3.167 | 0.0009 |
| 2.0 | 3.158 | 3.167 | 0.0009 |
| 2.2 | 3.158 | 3.167 | 0.0009 |
| 2.4 | 3.158 | 3.167 | 0.0009 |
| 2.6 | 2.892 | 2.899 | 0.007 |
| 2.8 | 2.653 | 2.656 | 0.003 |
| 3.0 | 2.653 | 2.656 | 0.003 |
| 3.2 | 2.653 | 2.656 | 0.003 |
| 3.4 | 2.653 | 2.656 | 0.003 |
| 3.6 | 2.653 | 2.656 | 0.003 |
| 3.8 | 2.653 | 2.656 | 0.003 |
| 4.0 | 2.288 | 2.305 | 0.017 |
| 4.2 | 2.288 | 2.305 | 0.017 |
| 4.4 | 2.183 | 2.191 | 0.008 |
| 4.6 | 1.969 | 1.981 | 0.012 |
| 4.8 | 1.969 | 1.981 | 0.012 |

the box on each link represents the active time slots for this link. Note that primary nodes $P_7, P_9$ and $P_{13}$ are helping relay secondary sessions' data while secondary nodes $S_1, S_3, S_{10}$ and $S_{15}$ are

Table 5.5: Performance comparison between the UPS policy and the interweave paradigms for different primary session rate requirements.

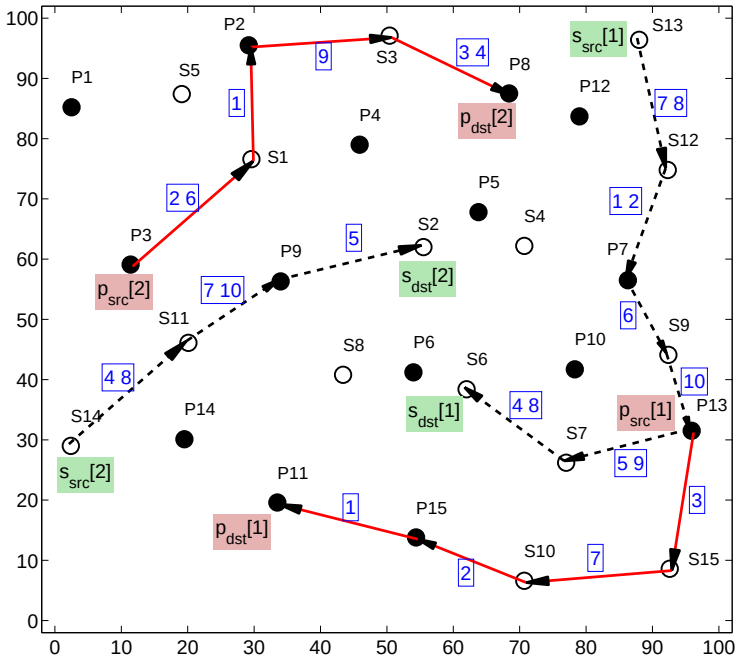
| | Rate Requirement | UPS | | Interweave Paradigm | |
|---|---|---|---|---|---|
| | $\hat{R}(1), \hat{R}(2)$ | Feasible in PN | SS Utility | Feasible in PN | SS Utility |
| Region 1 | 0 | Yes | 3.7012 | Yes | 3.0402 |
| | 0.2 | Yes | 3.288 | Yes | 1.899 |
| | 0.4 | Yes | 3.288 | Yes | 1.899 |
| | 0.6 | Yes | 3.288 | Yes | 1.899 |
| | 0.8 | Yes | 3.288 | Yes | 1.899 |
| | 1.0 | Yes | 3.288 | Yes | 1.899 |
| | 1.2 | Yes | 3.288 | Yes | 1.263 |
| | 1.4 | Yes | 3.288 | Yes | 1.263 |
| | 1.6 | Yes | 3.288 | Yes | 1.263 |
| | 1.8 | Yes | 3.158 | Yes | 1.425 |
| Region 2 | 2.0 | Yes | 3.158 | Yes | $-\infty$ |
| | 2.2 | Yes | 3.158 | Yes | $-\infty$ |
| | 2.4 | Yes | 3.158 | Yes | $-\infty$ |
| | 2.6 | Yes | 2.892 | Yes | $-\infty$ |
| | 2.8 | Yes | 2.653 | Yes | $-\infty$ |
| | 3.0 | Yes | 2.653 | Yes | $-\infty$ |
| | 3.2 | Yes | 2.653 | Yes | $-\infty$ |
| | 3.4 | Yes | 2.653 | Yes | $-\infty$ |
| | 3.6 | Yes | 2.653 | Yes | $-\infty$ |
| | 3.8 | Yes | 2.653 | Yes | $-\infty$ |
| Region 3 | 4.0 | Yes | 2.288 | No | N/A |
| | 4.2 | Yes | 2.288 | No | N/A |
| | 4.4 | Yes | 2.183 | No | N/A |
| | 4.6 | Yes | 1.969 | No | N/A |
| | 4.8 | Yes | 1.969 | No | N/A |
| Region 4 | 5.0 | Yes | $-\infty$ | No | N/A |
| | 5.2 | Yes | $-\infty$ | No | N/A |
| | 5.4 | Yes | $-\infty$ | No | N/A |
| | 5.6 | Yes | $-\infty$ | No | N/A |
| | 5.8 | Yes | $-\infty$ | No | N/A |
| | 6.0 | Yes | $-\infty$ | No | N/A |
| | 6.2 | Yes | $-\infty$ | No | N/A |
| | 6.4 | Yes | $-\infty$ | No | N/A |
| | 6.6 | Yes | $-\infty$ | No | N/A |
| | 6.8 | Yes | $-\infty$ | No | N/A |
| Region 5 | 7.0 | No | N/A | No | N/A |

helping relay the primary sessions' data. In comparison, under the interweave paradigm, the flow routing and scheduling for the primary network is shown in Fig. 5.4(b). According to the time slots used by the primary network, the secondary network calculates the remaining time slots at each node and uses them to maximize their rate utilities. The flow routing and scheduling for the secondary sessions under the interweave paradigm are also shown in Fig. 5.4(b). As expected, there is no cooperation at the node level between the two networks in terms of relaying each other's data.

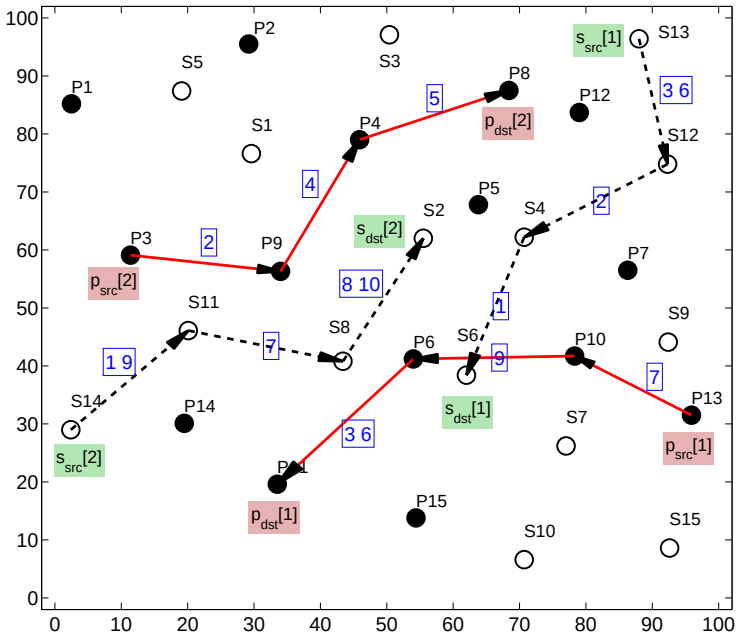
**Region 2** This region represents the scenario where the rate requirements of the primary sessions can be supported under both the UPS policy and the interweave paradigm, while the secondary sessions can only be supported under the UPS policy but not under the interweave paradigm (with zero rate for some sessions and thus $-\infty$ rate utility). This region contains that the combined network can offer more to the secondary sessions than the isolated networks under the interweave paradigm.

As an example, consider the case when the two primary sessions have rate requirements 3.0. The utility achieved for the secondary sessions under the UPS policy is 2.653. Specifically, the rates for the two secondary sessions are 3.753 and 2.793, respectively. Under the UPS policy, the flow routing and scheduling for the primary and secondary sessions are shown in Fig. 5.5(a). Note that primary nodes $P_7, P_9,$ and $P_{10}$ are helping relay secondary sessions' data while secondary nodes $S_1, S_3, S_7, S_{10}$ and $S_{15}$ are helping relay the primary sessions' data. Under the interweave paradigm, the flow routing and scheduling for primary network are shown in Fig. 5.5(b). Based on the time slots used by the primary network, the remaining time slots are not enough to support the secondary sessions, resulting in at least one of the secondary sessions with zero rate. Therefore, the rate utility for the secondary sessions is $-\infty$ under the interweave paradigm.

**Region 3** This region represents the scenario where the rate requirements of the primary sessions can be supported under the UPS policy but not so under the interweave paradigm. For the secondary sessions, there is still remaining resource to support them under the UPS policy. For fairness in comparison, we do not consider the rate utilities of the secondary sessions under the interweave paradigm (marked as "N/A"). Region 3 shows the definitive advantage of using a com-

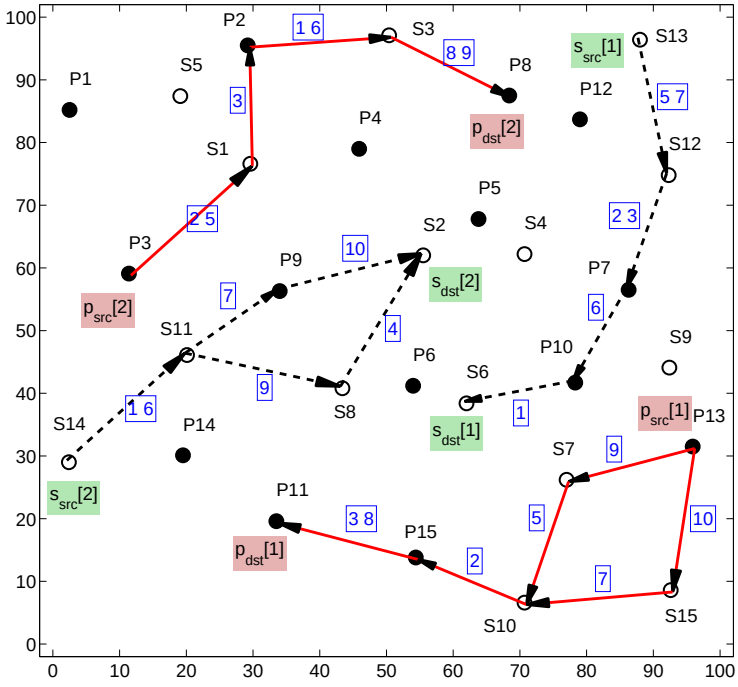



(a) UPS Policy

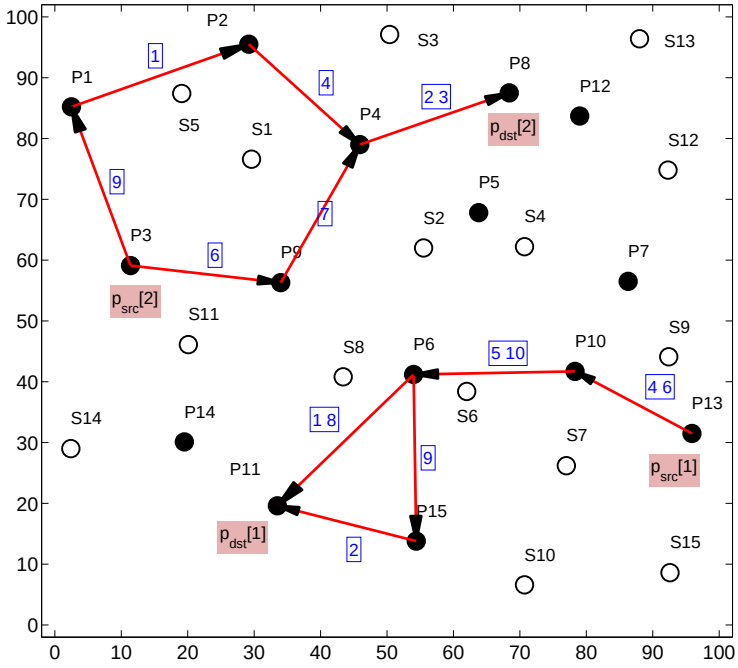



(b) Interweave Paradigm

Figure 5.4: A Region 1 example that showing the flow routing topologies and scheduling for the primary and secondary sessions, where the solid line segments are for the primary sessions while the dashed line segments are for the secondary sessions.




(a) UPS Policy




(b) Interweave Paradigm

Figure 5.5: A Region 2 example that showing the flow routing topologies and scheduling for the primary and secondary sessions.

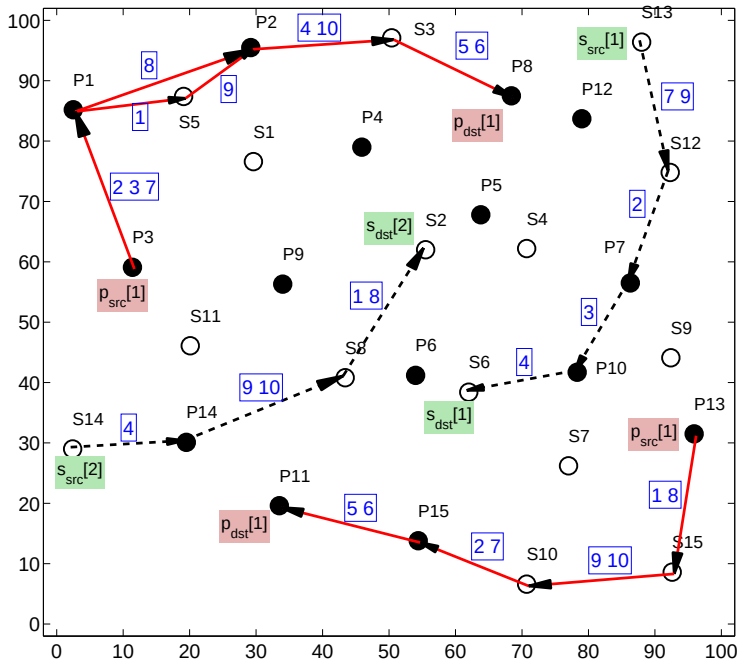



Figure 5.6: A Region 3 example that showing the flow routing topologies and scheduling for the primary and secondary sessions in the UPS policy.

bined network from the primary sessions' perspective over the interweave paradigm.

As an example, we consider the case when the two primary sessions have rate requirements 4.2. The utility objectives achieved by secondary sessions are 2.288 under the UPS policy. Specially, the rates for the two secondary sessions are 3.047 and 3.289, respectively. Under the UPS policy, the flow routing and scheduling for primary and secondary sessions are shown in Fig. 5.6. Note that primary nodes $P_7$ and $P_{14}$ are helping relay secondary sessions' data while secondary nodes $S_3, S_5, S_{10}$ and $S_{15}$ are helping relay the primary session' data.

**Region 4** This region represents the scenario where the rate requirement of the primary sessions can be satisfied under the UPS policy but not so under the interweave paradigm. The secondary sessions can no longer be supported under the UPS policy (with zero rate for at least one session and thus $-\infty$ rate utility). For fairness in comparison, we do not consider the rate utilities of the secondary sessions under the interweave paradigm (marked as "N/A") as even the rate requirements for the primary sessions cannot be supported. Similar to Region 3, this region shows the advantage

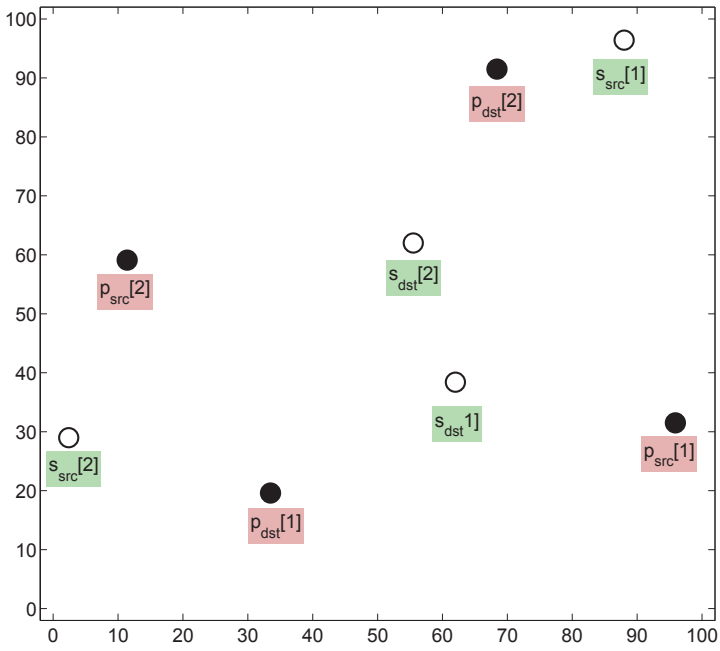



Figure 5.7: The locations of the source and destination nodes of the primary and secondary sessions.

of using a combined network to support the primary sessions over the interweave paradigm

**Region 5** As the rate requirements of the primary sessions continue to increase, even the UPS policy will no longer be able to support them after certain point. This is shown in Region 5.

### 5.6.3 Varying the Number of Nodes

In this section, we assume there are two primary sessions in the primary network and two secondary sessions in the secondary network. We fix the locations of source and destination nodes of the primary and secondary sessions as shown in Fig. 5.7. Then, we increase the number of primary and secondary nodes ($K$) in the network, and these nodes are uniformly deployed in the $100 \times 100$ area. Since these additional primary and secondary nodes only serve as relay nodes under UPS, there are no distinction between the two types of nodes.

Table 5.6 shows the average SS utility objective (over 100 network instances) under different

Table 5.6: The average SS utility objectives for different $K$ users.

| User Number $K$ | SS utility objectives |
|---|---|
| 5 | $-\infty$ |
| 10 | $-\infty$ |
| 15 | 1.9293 |
| 20 | 2.6653 |
| 25 | 3.2231 |
| 30 | 3.4772 |

number of nodes $(K)$ in the primary and secondary networks for the case when $\hat{R}(1) = 1.0$ and $\hat{R}(2) = 1.0$. When $K = 5$ and $K = 10$, the network is not dense enough and is not entirely connected. Therefore, the SS utility objectives are both $-\infty$ (i.e., the achievable secondary sessions rate is 0 in both cases). When $K = 15, 20, 25, 30$, the average SS utility objectives increase with the number of users $K$.

Then we vary $\hat{R}(1)$ and $\hat{R}(2)$ under different network size $K$. Fig. 5.8 shows the SS utility objectives under different network size $K$ when $\hat{R}(1)$ and $\hat{R}(2)$ vary. Again, for a given rate for $\hat{R}(1)$ and $\hat{R}(2)$, we have higher SS utility objectives under larger values of $K$.

### 5.6.4 Varying Session Numbers

In this section, we vary the primary and secondary session numbers. We randomly generate a 20-node primary network and a 20-node secondary network as shown in Fig. 5.9. In the first part, we will keep the number of secondary sessions fixed and vary the number of primary sessions. In the second part, we will do the converse, i.e., keep the number of primary sessions fixed and varying the number of secondary sessions. In both parts, we will compare the performance under the UPS policy and the interweave paradigm.

**Varying the Number of Primary Sessions.** Suppose that there are two secondary sessions,

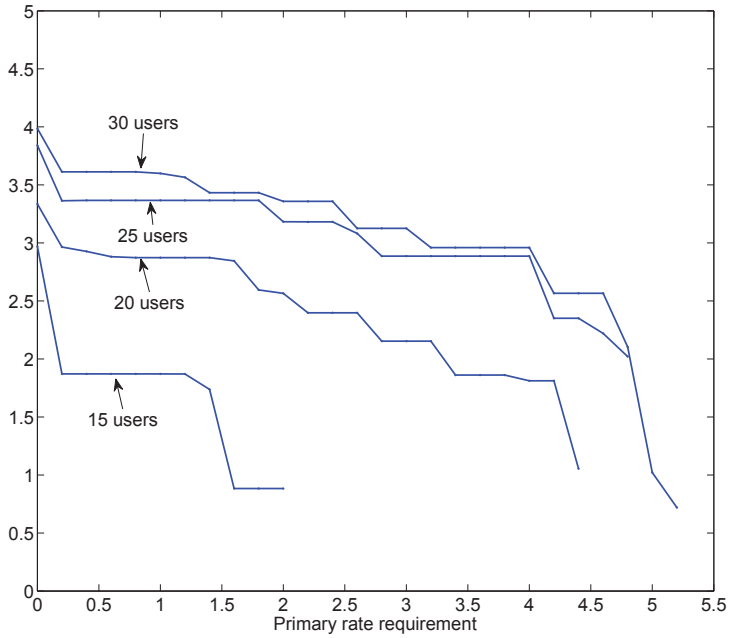



Figure 5.8: The comparison of the SS utility objectives for different number of nodes ($K = 10, 15, 20, 25,$ and $30$) with the increasing rate requirements for the primary sessions.

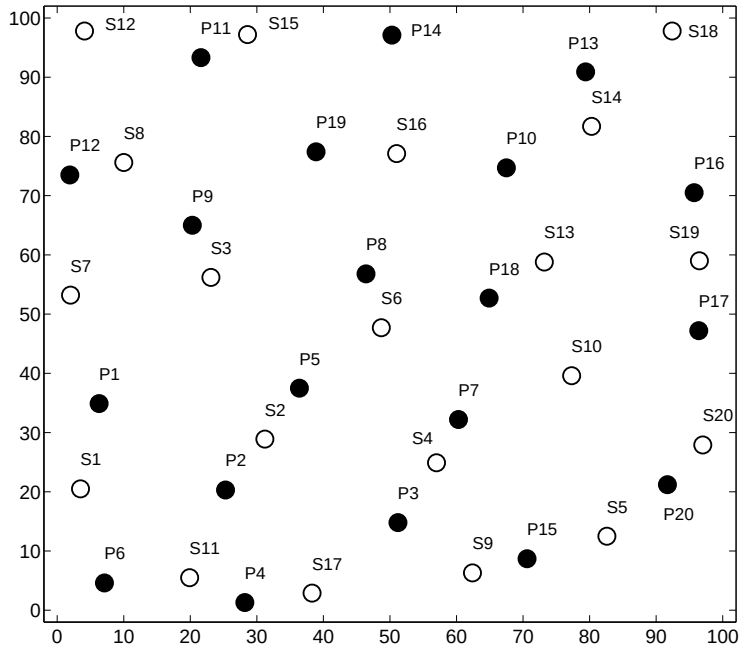



Figure 5.9: A 20-node primary network and a 20-node secondary network.

Table 5.7: Feasibility performance of the primary sessions and utilities of the secondary sessions under increasing number of the primary sessions.

| Number Of Primary Session | UPS | | Interweave Paradigm | |
|---|---|---|---|---|
| | Feasible in PN | Secondary Utility | Feasible in PN | Secondary Utility |
| 0 | Yes | 3.69 | Yes | 2.219 |
| 1 | Yes | 3.446 | Yes | 1.693 |
| 2 | Yes | 3.058 | Yes | 0.931 |
| 3 | Yes | 2.661 | Yes | 0.805 |
| 4 | Yes | 2.118 | Yes | $-\infty$ |
| 5 | Yes | 0.83 | Yes | $-\infty$ |
| 6 | Yes | -$\infty$ | No | N/A |
| 7 | Yes | -$\infty$ | No | N/A |
| 8 | No | N/A | No | N/A |

with each session's source and destination nodes being $(S_{11}, S_7)$ and $(S_4, S_1)$, respectively. By keeping these secondary sessions fixed, we increase the number of primary sessions. The source and destination nodes of each additional primary session is randomly chosen from the remaining primary nodes. Once chosen, we assume it has a data rate requirement of 1.8 and is added on top of the existing primary sessions. Table 5.7 shows our results. The first column in the table shows the increasing number of the primary sessions. The second and fourth columns show whether the additional new primary session can be accommodated (feasible) under UPS and interweave, respectively. Comparing these two columns, we can find that the maximum number of the primary sessions under UPS (7) is larger than that under interweave (5). The third and fifth columns show the utility function of the secondary sessions under UPS and interweave. Comparing these two columns, we can see that UPS achieves higher utility objectives than interweave. In summary, both primary and secondary sessions benefit more from UPS than interweave.

Table 5.8: Secondary sessions' utility values under increasing number of the secondary sessions.

| Number of Secondary Session | UPS | Interweave |
|---|---|---|
| 1 | 2.228 | 1.355 |
| 2 | 3.21 | 2.852 |
| 3 | 4.594 | 2.652 |
| 4 | 4.738 | 0.943 |
| 5 | 4.253 | $-\infty$ |
| 6 | 2.418 | $-\infty$ |
| 7 | 2.307 | $-\infty$ |
| 8 | 1.134 | $-\infty$ |
| 9 | $-\infty$ | $-\infty$ |

**Varying the Number of Secondary Sessions.** Now we do the converse. Suppose there are two primary sessions, with each session's source and destination nodes being $(P_9, P_{17})$ and $(P_1, P_{15})$, respectively. The data rate requirement for each primary session is 1.8. By keeping these primary sessions fixed, we increase the number of secondary sessions. The source and destination nodes of each additional secondary session is randomly chosen from the remaining secondary nodes. Once chosen, we add it on top of the existing secondary sessions. Table 5.8 shows our results. The first column in the table shows the increasing number of secondary sessions. The second and third columns show the utility values of the secondary sessions under UPS and interweave, respectively. Comparing these two columns, we can find that the maximum number of the secondary sessions that can be supported under UPS (8) is larger than that under interweave (4). Further, for the same number of secondary sessions (from 1 to 8), the achieved utility value under UPS is higher than that under interweave.

## 5.7 Chapter Summary

In this chapter, we develop a policy-based network cooperation paradigm as a new dimension for spectrum sharing between the primary and secondary users. Such network cooperation can be defined as a set of policies under which different degrees of cooperation are to be achieved. The benefits of this paradigm are numerous, including improved network connectivity and spatial diversity, increased flexibility in scheduling and routing, cost savings in infrastructure needed for each individual network, among others. For the purpose of performance study, we consider a specific policy called UPS, which allows a complete cooperation between the primary and secondary networks at the node level to relay each other's traffic. We study a problem with the goal of supporting the rate requirement of the primary network traffic while maximizing the throughput of the secondary sessions. Through rigorous mathematical modeling, problem formulation, approximation solution, and simulation results, we show that the UPS offers significantly better throughput performance than that under the interweave paradigm.

# Chapter 6

# Policy-based Network Cooperation: Throughput Region

## 6.1 Introduction

Recent push by the government agencies to share federal government radio spectrum with non-government entities has fueled the development of innovative technologies for spectrum sharing [46]. Coexistence of a secondary network with the primary network is the key to improve radio spectrum utilization. There has been extensive research on exploring coexistence between the primary and secondary networks in recent years. In [22], Goldsmith *et al.* outlined three coexistence paradigms, namely *interweave, underlay,* and *overlay.* These three paradigms were defined from an *information theoretic* perspective, solely based on how much side information (e.g., channel conditions, codebooks) is available to the secondary users. In the networking community, these three paradigms have been mapped into specific scenarios of how primary and secondary networks interact with each other for data forwarding. Specifically, the interweave paradigm refers to the simple idea that the secondary users are allowed to use a spectrum band allocated to the primary users only when the primary users are not using the band [3, 21, 72, 89]. This is the simplest approach to meet the current FCC requirements, which mandate that secondary users shall

not produce interference that is harmful to the primary users. This paradigm is analogous to the classic interference avoidance in medium access, or in cognitive radio (CR) terminology, *dynamic spectrum access* (DSA). This is the prevailing scenario to which most of research efforts have been devoted by the CR research community in recent years. The *underlay* paradigm refers to that secondary users' activities or interference on primary users is negligible (or below a given threshold). In contrast to the interweave paradigm, secondary users may be active *concurrently* with the primary users in the same area and in the same channel as long as the interference produced by the secondary nodes are controlled below a certain threshold (e.g., noise level). This can be achieved through a systematic interference cancelation (IC) by the secondary nodes without noticeable impact on the primary nodes [5, 23, 33, 75, 85, 86]. Finally, the overlay paradigm refers to having the secondary users offer some levels of cooperation with the primary users in data forwarding [28, 31, 42, 43, 56, 61, 83].

Under the interweave and underlay paradigms, the primary network would not feel the presence of the secondary network. The primary and secondary networks are independent in terms of data forwarding in each network. However, under the overlay paradigm, there is a certain level of cooperation on the data plane by the secondary network. Inspired by this cooperation idea in the overlay paradigm, there have been some recent efforts on how to exploit possible cooperation from the secondary users to help forward data for the primary users [28, 31, 42, 43, 56, 61, 83]. So far, these efforts have been limited to only having the secondary nodes help relay primary users' traffic. There is no consideration of the converse (i.e., primary users helping the secondary users), or a broader vision of a policy-based cooperation between the two networks. Such a limitation is mainly due to current FCC rules on existing wireless services and applications.

Recently, we proposed a novel policy-based cooperation between the primary and secondary networks [79]. We proposed to employ the *node-level (data plane) cooperation* as a new dimension for spectrum sharing between the primary and secondary users. Such network cooperation can be defined by a set of *policies* under which different degrees of cooperation can be achieved. Corresponding to each cooperation policy, a traffic-forwarding behavior for the primary and secondary users can be defined. A primitive policy used in [28, 31, 42, 43, 56, 61, 83] is to have the secondary

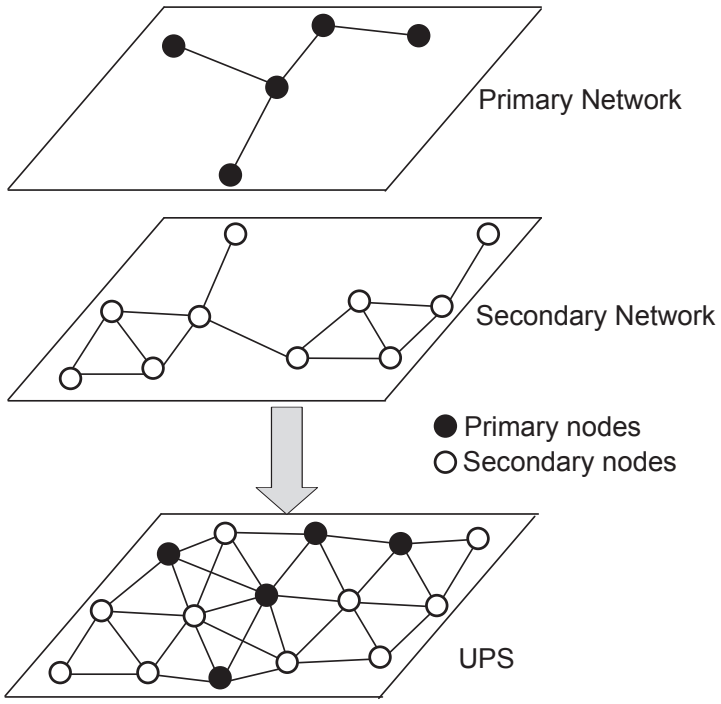



Figure 6.1: An illustration of the UPS policy for multi-hop primary and secondary networks.

network help relay primary users' traffic. Another policy, called UPS [79], is to allow complete node-level cooperation between the primary and secondary networks for data forwarding. These two examples are among many possible policies that one can define to achieve network sharing between the primary and secondary networks.

Figure 6.1 illustrates the UPS policy for a multi-hop primary and secondary network. It allows complete cooperation between the two networks on the data plane to help relay each other's traffic. Unlike the primitive policy, which is limited to only allowing secondary nodes help relay primary nodes' traffic, UPS allows primary nodes to help relay secondary nodes' traffic. From a network infrastructure perspective, the UPS policy allows to pool all the resources from both the primary and secondary networks together so that users in each network can access a much richer network infrastructure in a combined network. Note that although the two networks are combined into one at the node level, priority or service guarantee to the primary network traffic can still be enforced by implementing appropriate traffic engineering objectives. There are many poten-

tial benefits for node-level (data plane) cooperation between the primary and secondary networks. From the network perspective, the improved network connectivity, increased flexibility in power control, scheduling and routing all translate into improved forwarding performance for primary and secondary users' traffic. From a spectrum-sharing perspective, the ability to access other network infrastructure helps improve spatial diversity, thus allowing users to tap unused resources in the spatial domain. From an economic perspective, node-level sharing reduces the cost of building independent infrastructures for the primary and secondary networks. The UPS policy is ahead of today's FCC policies. But the benefits it offers may justify and propel it to become a viable approach for cooperation between the primary and secondary networks.

In this chapter, we offer an in-depth study of throughput performance for the UPS policy. In [79], we studied the maximum throughput for the secondary users while guaranteeing the rate requirement of the primary users. The proposed solution allows us to find a single point (a pair of throughput values for the primary and secondary users). Such a single-point solution does not offer a global view on the achievable throughput region between the primary and secondary users. In this chapter, we are interested in exploring the throughput region for users in the primary and secondary networks. Such a region (area) is bounded by the optimal throughput curve, which gives the maximum achievable throughput for users in the secondary (primary) network for *any* user throughput requirement in the primary (secondary) network. In other words, instead of having a single-point solution, the optimal throughput curve offers the entire landscape of maximum achievable throughput for both the primary and secondary users. Such a curve cannot be constructed by connecting discrete points found by the single-point solution approach in [79] since such an approach does not offer any optimality guarantee of the connected curve. As a result, a new solution method must be developed to find the optimal throughput curve.

The main contribution of this chapter is the development of a solution to find the optimal throughput curve and thus, the throughput region under the curve. To do this, we formulate a *multiobjective* optimization problem that maximizes the throughput for both the primary and secondary users. We show how to transform this multiobjective problem into a single-objective problem by using a novel approach called *weighted Chebyshev norm* [19, 50, 66]. For the transformed

single-objective program, we exploit its mathematical structure, and propose a method to find new Pareto-optimal points iteratively. When the total number of Pareto-optimal points is not large, our algorithm can find all Pareto-optimal points. The throughput curve obtained from these Pareto-optimal points will be the exact optimal throughput curve. When the number of Pareto-optimal points is large, we present a termination condition upon which the final throughput curve is $\varepsilon$-approximation to the optimal curve meaning that the approximation error is no more than $\varepsilon$ – a predefined approximation error. We conduct a case study to demonstrate how to find all Pareto-optimal points iteratively and how to construct the optimal throughput curve by our algorithm. For each point on this throughput curve, we show how to construct a feasible solution based on the solution of its corresponding Pareto-optimal point. By comparing the throughput region (the area under the throughput curve) between the UPS policy and interweave, we show that the throughput region under the UPS policy is substantially larger than that under interweave.

The remainder of this chapter is organized as follows. In Section 6.2, we present the mathematical model for the primary and secondary networks with node-level cooperation. We also present a multiobjective formulation for maximizing both the primary and secondary users' throughput. In Section 6.3, we develop an efficient algorithm to find $\varepsilon$-approximation to the optimal throughput curve for the throughput region. Section 6.4 presents results for a case study and Section 6.5 concludes this chapter.

## 6.2 Mathematical Modeling and Formulation

### 6.2.1 Network Model

We consider a multi-hop secondary network co-located with a multi-hop primary network. Denote the set of primary nodes as $\hat{\mathcal{N}}_{\mathrm{P}}$ and the set of secondary nodes as $\mathcal{N}_{\mathrm{S}}$. Denote $\mathcal{N}$ as the combined set of nodes from both networks, i.e., $\mathcal{N} = \hat{\mathcal{N}}_{\mathrm{P}} \cup \mathcal{N}_{\mathrm{S}}$. We assume there is a set of channels $\mathcal{B}$ available in the primary network. Suppose that there are $T$ time slots in each time frame. Denote $\hat{\mathcal{L}}$ and $\mathcal{L}$ as the set of primary and secondary user sessions, respectively. For each primary session

Table 6.1: Notation

| | |
|---|---|
| Primary Network | |
| $\hat{\mathcal{N}}_{\mathrm{P}}$ | The set of primary nodes |
| $\hat{\mathcal{L}}$ | The set of primary sessions |
| $\hat{f}_{ij}(l)$ | The flow rate traversing on link $(i,j)$ that attributes to primary session $l \in \hat{\mathcal{L}}$ |
| $\hat{s}(l)$ | The source node of primary session $l \in \hat{\mathcal{L}}$ |
| $\hat{d}(l)$ | The destination node of primary session $l \in \hat{\mathcal{L}}$ |
| $\hat{r}(l)$ | The data rate achieved by primary session $l \in \hat{\mathcal{L}}$ |
| $\hat{r}_{\min}$ | The minimum data rate among all primary sessions |
| Secondary Network | |
| $\mathcal{N}_{\mathrm{S}}$ | The set of secondary nodes |
| $\mathcal{L}$ | The set of secondary sessions |
| $f_{ij}(m)$ | The flow rate traversing on link $(i,j)$ that is attributed to secondary session $m \in \mathcal{L}$ |
| $s(m)$ | The source node of secondary session $m \in \mathcal{L}$ |
| $d(m)$ | The destination node of secondary session $m \in \mathcal{L}$ |
| $r(m)$ | The data rate achieved by secondary session $m \in \mathcal{L}$ |
| $r_{\min}$ | The minimum data rate among all secondary sessions. |
| Combined Network | |
| $\mathcal{N}$ | The set of all nodes in the network, $\mathcal{N} = \hat{\mathcal{N}}_{\mathrm{P}} \bigcup \mathcal{N}_{\mathrm{S}}$ |
| $C_{ij}$ | The link capacity of link $(i,j), i,j \in \mathcal{N}$ |
| $x_{ij}^{b}(t)$ | $= 1$ if node $i$ is transmitting data to node $j$ in time slot $t$ |
| | on channel $b$, and is $0$ otherwise |
| $\mathcal{T}_i$ | The set of nodes that are located within the transmission range of node $i \in \mathcal{N}$ |
| $\mathcal{I}_i$ | The set of nodes that are located within the interference range of node $i \in \mathcal{N}$ |
| $T$ | The number of time slots in a frame |
| $\mathcal{B}$ | The available channels in the network |

$l \in \hat{\mathcal{L}}$, denote $\hat{r}(l)$ as the data rate of this session $l$. Likewise, for each secondary session $m \in \mathcal{L}$, denote $r(m)$ as the data rate of this session $m$. The primary and secondary networks are allowed to share their nodes, in addition to channels. The goal of this chapter is to find the optimal throughput curve for users in the primary and secondary networks. In contrast, the objective in [79] is only to maximize secondary network throughput with a fixed primary network throughput requirement. In other words, the solution in [79] only constitutes a single-point in the optimal throughput curve in this chapter. Table 6.1 lists notation in this chapter.

### 6.2.2 Interference Modeling

In the combined network, denote $\mathcal{T}_i$ as the set of nodes in $\mathcal{N}$ (including both the primary and secondary nodes) that is located within node $i$'s transmission range, where $i$ can be either a primary or secondary node (i.e., $i \in \mathcal{N}$). Denote $\mathcal{I}_j$ as the set of nodes in $\mathcal{N}$ (including both primary and secondary nodes) that is located within node $j$'s interference range, where $j$ can be either a primary or secondary node (i.e., $j \in \mathcal{N}$).

**Self-interference Constraints.** We assume scheduling is done on both channels and time slots. We use a binary variable $x_{ij}^b(t), i, j \in \mathcal{N}, b \in \mathcal{B}$ and $1 \leq t \leq T$, to indicate whether node $i$ transmits data to node $j$ on channel $b$ in time slot $t$. That is,

$$x_{ij}^b(t) = \begin{cases} 1 & \text{If node } i \text{ transmits data to node } j \text{ on channel } b \text{ in time slot } t; \\ 0 & \text{otherwise.} \end{cases}$$

where $i \in \mathcal{N}, j \in \mathcal{T}_i, b \in \mathcal{B}$, and $1 \leq t \leq T$.

Since each primary or secondary session is unicast, node $i$ only needs to transmit to or receive from one node on a channel and in a time slot. We have:

$$\sum_{j \in \mathcal{T}_i} x_{ij}^b(t) \leq 1 \qquad (i \in \mathcal{N}, b \in \mathcal{B}, 1 \leq t \leq T)\ , \tag{6.2.1}$$

$$\sum_{k \in \mathcal{T}_i} x_{ki}^b(t) \leq 1 \qquad (i \in \mathcal{N}, b \in \mathcal{B}, 1 \leq t \leq T)\ . \tag{6.2.2}$$

Assuming half-duplex at each node $i$, then we have:

$$x_{ij}^b(t) + x_{ki}^b(t) \leq 1 \qquad (i \in \mathcal{N}, j, k \in \mathcal{T}_i, b \in \mathcal{B}, 1 \leq t \leq T) \ . \tag{6.2.3}$$

The three constraints in (6.2.1), (6.2.2) and (6.2.3) can be replaced by the following single and equivalent constraint:

$$\sum_{j \in \mathcal{T}_i} x_{ij}^b(t) + \sum_{k \in \mathcal{T}_i} x_{ki}^b(t) \leq 1 \qquad (i \in \mathcal{N}, b \in \mathcal{B}, 1 \leq t \leq T). \tag{6.2.4}$$

**Mutual Interference Constraints.** For any primary or secondary node $j \in \mathcal{N}$ that is receiving data on channel $b$ in time slot $t$, it shall not be interfered by another (unintended) transmitting node $p \in \mathcal{I}_j$ on the same channel and time slot. We have the following mutual interference constraint:

$$x_{ij}^b(t) + x_{pk}^b(t) \leq 1 \ , \tag{6.2.5}$$

where $i \in \mathcal{T}_j, p \in \mathcal{I}_j, k \in \mathcal{T}_p, j \in \mathcal{N}, j \neq k, b \in \mathcal{B}$ and $1 \leq t \leq T$.

Following the same token in (6.2.4), the three constraints in (6.2.1), (6.2.2) and (6.2.5) can be replaced by the following single and equivalent constraint:

$$\sum_{i \in \mathcal{T}_j} x_{ij}^b(t) + \sum_{k \in \mathcal{T}_p} x_{pk}^b(t) \leq 1 \ , \tag{6.2.6}$$

where $p \in \mathcal{I}_j, j \in \mathcal{N}, j \neq k, b \in \mathcal{B}$ and $1 \leq t \leq T$.

### 6.2.3 Traffic Modeling

**Flow Routing for Primary Sessions.** For flexibility and load balancing, we allow flow splitting in the network. That is, the flow rate of a session may split and merge inside the network $\mathcal{N}$ in whatever loop-free manner as long as it can maximize the data rate $\hat{r}(l)$ of session $l \in \hat{\mathcal{L}}$. Denote $\hat{s}(l)$ and $\hat{d}(l)$ as the source and destination nodes of primary session $l \in \hat{\mathcal{L}}$, respectively. Denote $\hat{f}_{ij}(l)$ as the data rate on link $(i, j)$ that is attributed to primary session $l \in \hat{\mathcal{L}}$, where $i \in \mathcal{N}$ and $j \in \mathcal{T}_i$. We have the following flow balance constraints:

- If node $i$ is the source node of primary session $l \in \hat{\mathcal{L}}$ (i.e., $i = \hat{s}(l)$), then

$$\sum_{j \in \mathcal{T}_i} \hat{f}_{ij}(l) = \hat{r}(l) \qquad (l \in \hat{\mathcal{L}}, i = \hat{s}(l)). \tag{6.2.7}$$

- If node $i$ is an intermediate relay node for primary session $l$ (i.e., $i \neq \hat{s}(l)$ and $i \neq \hat{d}(l)$), then

$$\sum_{j \in \mathcal{T}_i}^{j \neq \hat{s}(l)} \hat{f}_{ij}(l) = \sum_{k \in \mathcal{T}_i}^{k \neq \hat{d}(l)} \hat{f}_{ki}(l) \qquad (l \in \hat{\mathcal{L}}, i \in \mathcal{N}). \tag{6.2.8}$$

- If node $i$ is the destination node of primary session $l$ (i.e., $i = \hat{d}(l)$), then

$$\sum_{k \in \mathcal{T}_i} \hat{f}_{ki}(l) = \hat{r}(l) \qquad (l \in \hat{\mathcal{L}}, i = \hat{d}(l)). \tag{6.2.9}$$

It can be easily verified that once (6.2.7) and (6.2.8) are satisfied, then (6.2.9) is also satisfied. As a result, it is sufficient to list only (6.2.7) and (6.2.8) in the formulation.

**Flow Routing for Secondary Sessions.** Denote $s(m)$ and $d(m)$ as the source and destination nodes of secondary session $m \in \mathcal{L}$, respectively. Denote $f_{ij}(m)$ as the data rate on link $(i, j)$ that is attributed to secondary session $m \in \mathcal{L}$. Similar to that for the primary sessions, we allow flow splitting for the secondary sessions. We have the following flow balance constraints:

- If node $i$ is the source node of secondary session $m \in \mathcal{L}$ (i.e., $i = s(m)$), then we have

$$\sum_{j \in \mathcal{T}_i} f_{ij}(m) = r(m) \qquad (m \in \mathcal{L}, i = s(m)) \tag{6.2.10}$$

- If node $i$ is an intermediate relay node for secondary session $m$ (i.e., $i \neq s(m)$ and $i \neq d(m)$), then

$$\sum_{j \in \mathcal{T}_i}^{j \neq s(m)} f_{ij}(m) = \sum_{k \in \mathcal{T}_i}^{k \neq d(m)} f_{ki}(m) \qquad (m \in \mathcal{L}, i \in \mathcal{N}), \tag{6.2.11}$$

- If node $i$ is the destination node of secondary session $m$ (i.e., $i = d(m)$), then

$$\sum_{k \in \mathcal{T}_i} f_{ki}(m) = r(m) \qquad (m \in \mathcal{L}, i = d(m)) \tag{6.2.12}$$

Again, to avoid redundancy, it is sufficient to list only (6.2.10) and (6.2.11) in the formulation.

**Link Capacity Constraints.** For each link $(i, j)$, denote the link capacity as $C_{ij}$, i.e., $C_{ij} = W \log_2(1 + \frac{\rho_i d_{ij}^{-\gamma} \lambda}{N_0})$, where $W$ is the bandwidth, $\rho_i$ is the power spectral density from transmit node $i$, $d_{ij}$ is the distance between nodes $i$ and $j$, $\gamma$ is the path loss index, $\lambda$ is the antenna related constant, and $N_0$ is the ambient Gaussian noise density. Since the aggregate flow rate from the primary and secondary sessions on each link $(i, j)$ cannot exceed the average link rate (over $T$ time slots), we have

$$\sum_{l \in \hat{\mathcal{L}}}^{j \neq \hat{s}(l), i \neq \hat{d}(l)} \hat{f}_{ij}(l) + \sum_{m \in \mathcal{L}}^{j \neq s(m), i \neq d(m)} f_{ij}(m) \leq \frac{1}{T} \sum_{t=1}^{T} \sum_{b \in \mathcal{B}} C_{ij} \cdot x_{ij}^{b}(t). \tag{6.2.13}$$

### 6.2.4 Multiobjective Formulation

Our goal is to find the optimal throughput curve for both the primary and secondary sessions. This problem can be formulated as a *multicriteria* optimization program with the objectives of maximizing session throughput in both primary and secondary networks. For throughput maximization, we maximize the minimum session rate in each network to ensure fairness. We define $\hat{r}_{\min}$ and $r_{\min}$ as the minimum rate among the primary and secondary sessions, respectively. Then we have:

$$\hat{r}_{\min} \leq \hat{r}(l) \qquad (l \in \hat{\mathcal{L}}) \,, \tag{6.2.14}$$

$$r_{\min} \leq r(m) \qquad (m \in \mathcal{L}) \,. \tag{6.2.15}$$

The multiobjective program can be written as follows:

BIOPT

max $\hat{r}_{\min}$

max $r_{\min}$

s.t. Self interference constraints: (6.2.4);

Mutual interference constraints:(6.2.6);

Flow routing for primary sessions: (6.2.7), (6.2.8);

Flow routing for secondary sessions: (6.2.10), (6.2.11);

Link capacity constraints: (6.2.13);

Minimum sessions rate constraints: (6.2.14), (6.2.15).

In this formulation, $C_{ij}$ are constants, $x_{ij}^b(t)$ are binary variables, $\hat{f}_{ij}(l)$, $f_{ij}(m)$ $\hat{r}(l)$, $\hat{r}_{\min}$, $r(m)$ and $r_{\min}$ are continuous variables. This formulation is in the form of *multiobjective mixed-integer linear programming* (MOMILP). In the next section, we develop an efficient algorithm to solve this problem.

## 6.3 An Approximation Algorithm

### 6.3.1 Background and Roadmap

For optimization problem BIOPT, we want to maximize the minimum achievable throughput in both the primary and secondary networks. Since the two objective functions, $\hat{r}_{\min}$ and $r_{\min}$, are conflicting with each other, we pursue *Pareto-optimal* solutions [20]. For ease of exposition, we define $\alpha = \{\mathbf{x}, \mathbf{f}, \hat{\mathbf{f}}, \mathbf{r}, \hat{\mathbf{r}}, r_{\min}, \hat{r}_{\min}\}$ as a feasible solution to BIOPT, where $\mathbf{x}, \mathbf{f}, \hat{\mathbf{f}}, \mathbf{r}$, and $\hat{\mathbf{r}}$ represent the set of $x_{ij}, f_{ij}, \hat{f}_{ij}, r(m)$ and $\hat{r}(l)$ for $i \in \mathcal{N}$, $j \in \mathcal{N}$, $l \in \hat{\mathcal{L}}$ and $m \in \mathcal{L}$. For a feasible solution $\alpha$, we denote $U(\alpha)$ and $V(\alpha)$ as

$$U(\alpha) = \hat{r}_{\min} \ , \tag{6.3.1}$$

$$V(\alpha) = r_{\min} \ . \tag{6.3.2}$$

Then BIOPT can be re-written as follows:

BIOPT

$$\begin{aligned} \max \quad & U(\alpha) \\ \max \quad & V(\alpha) \\ \text{s.t.} \quad & \alpha = \{\mathbf{x}, \mathbf{f}, \hat{\mathbf{f}}, \mathbf{r}, \hat{\mathbf{r}}, r_{\min}, \hat{r}_{\min}\}; \\ & \text{Constraints (6.2.4), (6.2.6)-(6.2.8), (6.2.10), (6.2.11), (6.2.13)-(6.3.2).} \end{aligned}$$

For a Pareto-optimal solution $\alpha^\dagger$, the corresponding objective pair $(U^\dagger, V^\dagger)$ is called a *Pareto-optimal point.* For a Pareto-optimal point $(U^\dagger, V^\dagger)$, there does not exist another feasible solution $\alpha$ with objective pair $(U, V)$ such that $U \geq U^\dagger$ and $V > V^\dagger$, or $U > U^\dagger$ and $V \geq V^\dagger$. This means that it is impossible to further improve any one objective without deteriorating the other. For our problem, it is difficulty to find Pareto-optimal point directly. Therefore, we find a *weakly Pareto-optimal point* first and then find the corresponding Pareto-optimal point. For a feasible solution $\alpha^*$, with corresponding objective pair $(U^*, V^*)$, if there does not exist any other solution $\alpha$ with its objective pair $(U, V)$ such that $U > U^*$ and $V > V^*$, then solution $\alpha^*$ is called a *weakly Pareto-optimal solution* and $(U^*, V^*)$ is called a *weakly Pareto-optimal point.* From this definition, it is obvious that a Pareto-optimal point is also a weakly Pareto-optimal point, while a weakly Pareto-optimal point is not always Pareto optimal.

To find all the Pareto-optimal points for BIOPT, we can combine the two objectives into a single criterion. There are two main techniques to transform a multiobjective problem into a single-objective problem: (i) weighted sum method and (ii) Chebyshev norm method. In the weighed sum method, the objective is defined as a nonnegative linear combination of the two objective functions through a parameter $0 \leq \beta \leq 1$:

$$\max \ \beta \cdot U(\alpha) + (1 - \beta) \cdot V(\alpha). \tag{6.3.3}$$

Although it is easy to find a Pareto-optimal point for a given $\beta$, it is difficult to find *all* Pareto-optimal points using this method. This is because there is an infinite number of $\beta$ values between $[0, 1]$ and it is impossible to check out all these values for Pareto-optimal points. So the weighted sum method is not a good choice to solve our problem.

In this chapter, we employ the Chebyshev norm method, which allows us to find all Pareto-optimal points by identifying *specific* values of $\beta$ (instead of enumerating all values blindly). The Chebyshev norm between two points $A$ and $B$ with $(U_A, V_A)$ and $(U_B, V_B)$, respectively, is defined as follows:

$$\|A - B\| = \max\{|U_A - U_B|, |V_A - V_B|\}. \tag{6.3.4}$$

The *weighted* Chebyshev norm with weight $0 \leq \beta \leq 1$ is defined as follows:

$$\|A - B\|_\beta = \max\{\beta|U_A - U_B|, (1-\beta)|V_A - V_B|\}. \tag{6.3.5}$$

In the rest of this section, we give the single-objective problem formulation from BIOPT via weighted Chebyshev norm. Then we show how to find new Pareto-optimal points by properly setting the value of $\beta$ in each iteration. In the case when there is an infinite number of Pareto-optimal points, we show how to terminate the iteration when we have achieved $\varepsilon$-approximation in the objective value. Finally, by connecting all Pareto-optimal points that we found in the iterations, we obtain the throughput curve and prove that its approximation error to optimal is no more than $\varepsilon$.

### 6.3.2 Single Objective Formulation with Chebyshev Norm

To transform our multiobjective problem into a single objective problem, we define an ideal point $I$ with coordinate $(U_I, V_I)$ such that for any feasible solution $\alpha$ with $(U(\alpha), V(\alpha))$, we have $U(\alpha) \leq U_I$ and $V(\alpha) \leq V_I$. In other words, $U_I$ is an upper bound of $U(\alpha)$ and $V_I$ is an upper bound of $V(\alpha)$, respectively, for any $\alpha$. Based on this ideal point $I$, we define weighted Chebyshev norm between a feasible solution point $\alpha$ with $(U(\alpha), V(\alpha))$ and $(U_I, V_I)$ as $\max\{\beta|U_I - U(\alpha)|, (1-\beta)|V_I - V(\alpha)|\}$. We are interested in the minimum value of weighted Chebyshev norm over all feasible solutions, i.e.,

$$\min_\alpha \max\{\beta|U_I - U(\alpha)|, (1-\beta)|V_I - V(\alpha)|\}, \tag{6.3.6}$$

where the minimization is taken over all feasible solutions $\alpha$ for BIOPT, and $\beta \in [0, 1]$. We now show that for a given $\beta$, the optimal objective pair(s) $(U(\alpha), V(\alpha))$ (may not be unique) in (6.3.6) are weakly Pareto-optimal points.

**Lemma 6.1.** For any given $\beta \in [0, 1]$, the optimal objective pairs from (6.3.6) are weakly Pareto-optimal points.

*Proof.* This proof is based on contradiction. For a given $\beta$, suppose that the optimal objective pairs $H$ with $(U_H, V_H)$ achieves the minimum Chebyshev norm for (6.3.6), but $H$ is not a weakly Pareto-optimal point. Then, there must exists another point $K$ with objective pair $(U_K, V_K)$, that has $U_K > U_H$ and $V_K > V_H$. The weighted Chebyshev norms between $H$ and $I$, $K$ and $I$ are $\max\{\beta(U_I - U_H), (1-\beta)(V_I - V_H)\}$ and $\max\{\beta(U_I - U_K), (1-\beta)(V_I - V_K)\}$, respectively. Since $(U_I - U_K) < (U_I - U_H)$ and $(V_I - V_K) < (V_I - V_H)$, we have $\beta(U_I - U_K) < \beta(U_I - U_H)$ and $(1-\beta)(V_I - V_K) < (1-\beta)(V_I - V_H)$. Therefore, $\max\{\beta(U_I - U_K), (1-\beta)(V_I - V_K)\} < \max\{\beta(U_I - U_H), (1-\beta)(V_I - V_H)\}$. This means $K$ can achieve a smaller Chebyshev norm than $H$, which contradicts with the assumption that $(U_H, V_H)$ can achieve the minimum Chebyshev norm. Therefore, for any given $\beta \in [0, 1]$, the optimal objective pair $H$ that achieves the minimum chebyshev norm for (6.3.6) is always weakly Pare-optimal point. □

There is an infinite number of points that can be used as the ideal point. For simplicity, we choose our ideal point $I$ with $(U_I, V_I)$ as follows. For $U_I$, we set it to the maximum objective value of $U$ when $V$ is set to 0 in BIOPT. Likewise, for $V_I$, we set it to the maximum objective value of $V$ when $U$ is set to 0 in BIOPT. Then, we have a single objective formulation as follows:

$$
\begin{aligned}
\min \quad & \max\{\beta|U_I - U(\alpha)|, (1-\beta)|V_I - V(\alpha)|\} \\
\text{s.t.} \quad & \alpha = \{\mathbf{x}, \mathbf{f}, \hat{\mathbf{f}}, \mathbf{r}, \hat{\mathbf{r}}, \mathbf{r}_{\min}, \hat{\mathbf{r}}_{\min}\}; \\
& \beta \in [0, 1]; \\
& \text{Constraints (6.2.4), (6.2.6)-(6.2.8), (6.2.10), (6.2.11), (6.2.13)-(6.3.2).}
\end{aligned}
$$

Since the objective function in the above formulation is nonlinear, we define $z = \max\{\beta|U_I - U(\alpha)|, (1-\beta)|V_I - V(\alpha)|\}$. Then we have:

$$
\begin{aligned}
\text{BIOPT-L } \min \quad & z \\
\text{s.t.} \quad & z \geq \beta(U_I - U(\alpha)); \\
& z \geq (1-\beta)(V_I - V(\alpha)); \\
& \alpha = \{\mathbf{x}, \mathbf{f}, \hat{\mathbf{f}}, \mathbf{r}, \hat{\mathbf{r}}, \mathbf{r}_{\min}, \hat{\mathbf{r}}_{\min}\}; \\
& \beta \in [0,1]; \\
& \text{Constraints (6.2.4), (6.2.6)-(6.2.8), (6.2.10), (6.2.11), (6.2.13)-(6.3.2).}
\end{aligned}
$$

Now, the objective function is linear. For a given $\beta$, BIOPT-L is in the form of mixed-integer linear program (MILP), which is NP-hard in general. But fortunately, all integer variables in this MILP are binary. For binary variables that can only take 0 and 1, a branch-and-cut based solution procedure used by a commerical solver such as CPLEX is very efficient. Therefore, we will use CPLEX to solve all our binary MILP problems, which turns out to be very successful for all practical purposes.

### 6.3.3 Finding Pareto-optimal Point for a Given $\beta$

From Lemma 6.1, we know that for a given $\beta$, the optimal objective pair obtained from BIOPT-L is a weakly Pareto-optimal point. For this weakly Pareto-optimal point $(U^*, V^*)$, we can find the corresponding Pareto-optimal point $(U^\dagger, V^\dagger)$ based on the following algorithm:

**Algorithm 6.1. (Weakly Pareto to Pareto)**

8. Input: Weakly Pareto-optimal point $(U^*, V^*)$.
9. Let $V = V^*$. Solve BIOPT to obtain the optimal $U^\dagger$.
10. Let $U = U^\dagger$. Solve BIOPT to obtain the optimal $V^\dagger$.
11. Return $(U^\dagger, V^\dagger)$.

From line 2 in Algorithm 6.1, we know there does not exist another point $(U, V)$ with $U > U^\dagger$ and $V \geq V^*$. From line 3, we know that $V^\dagger \geq V^*$, and there does not exist any other point with $U \geq U^\dagger$ and $V > V^\dagger$. Therefore, there does not exist any other point $(U, V)$ with $U > U^\dagger$ and

$V \geq V^{\dagger}$, or $U \geq U^{\dagger}$ and $V > V^{\dagger}$. Then $(U^{\dagger}, V^{\dagger})$ is a Pareto-optimal point. It is obvious that the weakly Pareto-optimal point $(U^{*}, V^{*})$ and the corresponding Pareto-optimal point $(U^{\dagger}, V^{\dagger})$ can achieve the same $z$ for BIOPT-L. We omit its proof here to conserve space.

### 6.3.4 Determination of New Pareto-optimal Points

In the last section, we showed that for a given $\beta$, we can find its corresponding Pareto-optimal point. Since there is an infinite number of values for $\beta$ in $[0, 1]$ and different $\beta$ values may correspond to the same Pareto-optimal point, it is important to identify a subset of $\beta$ values that allow us to find all Pareto-optimal points. In this chapter, we propose a method to determine the $\beta$ value based on two given Pareto-optimal points that allows us to find a new Pareto-optimal point.

In order to derive the representation for $\beta$ with any two known pare-optimal point, we start from one simple scenario. We assume that there are a total of three Pareto-optimal points, where the two extreme Pareto-optimal points $A$ with $(U_A, V_A)$ and $B$ with $(U_B, V_B)$ are known, and the Pareto-optimal point $K$ (with $(U_K, V_K)$) lies *strictly between $A$ and $B$* (or *is a new Pareto-optimal point that lies between $A$ and $B$*) defined as $U_A < U_K < U_B$ and $V_A > V_K > V_B$. The following lemma derives the representation of $\beta = \beta_{AB}$ that is necessary for generating such Pareto-optimal point $K$ based on $A$ and $B$.

**Lemma 6.2.** To generate any possible Pareto-optimal point $K$ that lies strictly between $A$ and $B$, $\beta_{AB}$ should be given by

$$\beta_{AB} = \frac{(V_I - V_B)}{(U_I - U_A + V_I - V_B)}. \tag{6.3.7}$$

*Proof.* For the new Pareto-optimal point $K$ between $A$ and $B$, we have $U_A < U_K < U_B$ and $V_A > V_K > V_B$. If we want to generate the new Pareto-optimal point $K$, we need to have $z_K < z_A$ and $z_K < z_B$. So we want to show that if $z_K < z_A$ and $z_K < z_B$ hold true, then $\beta_{AB}$ must be given by (6.3.7).

We first explore the necessary condition for $\beta_{AB}$ if $z_K < z_A$ always holds.

- If $\beta_{AB}(U_I - U_K) > (1 - \beta_{AB})(V_I - V_K)$, then $z_K = \beta_{AB}(U_I - U_K)$. Since $U_A < U_K$ and $V_A > V_K$, then we have:

$$z_A \geq \beta_{AB}(U_I - U_A) > \beta_{AB}(U_I - U_K) = z_K.$$

  Hence, if $z_K < z_A$ always holds, $\beta_{AB}$ can be any value in $(0, 1]$.

- If $\beta_{AB}(U_I - U_K) \leq (1 - \beta_{AB})(V_I - V_K)$, then $z_K = (1 - \beta_{AB})(V_I - V_K)$. Based on $U_A < U_K$ and $V_A > V_K$, we have the following constraints:

$$\beta_{AB}(U_I - U_K) < \beta_{AB}(U_I - U_A),$$
$$(1 - \beta_{AB})(V_I - V_K) > (1 - \beta_{AB})(V_I - V_A).$$

  Since $z_K = (1 - \beta_{AB})(V_I - V_K) > (1 - \beta_{AB})(V_I - V_A)$, and $z_A = \max\{\beta_{AB}(U_I - U_A), (1 - \beta_{AB})(V_I - V_A)\}$, then, $z_K < z_A$ implies that $(1 - \beta_{AB})(V_I - V_K) < \beta_{AB}(U_I - U_A)$. But this should hold for any $V_K$ satisfying $V_K > V_B$, and so, we must have $(1 - \beta_{AB})(V_I - V_B) \leq \beta_{AB}(U_I - U_A)$, i.e., $\beta_{AB}(U_I - U_A + V_I - V_B) \geq V_I - V_B$. Thus,

$$\beta_{AB} \geq \frac{V_I - V_B}{U_I - U_A + V_I - V_B}. \tag{6.3.8}$$

Therefore, we can conclude that if $z_K < z_A$ always holds, the necessary condition for $\beta_{AB}$ is given by (6.3.8).

Next, we explore the necessary condition for $\beta_{AB}$ if $z_K < z_B$ always holds.

- If $\beta_{AB}(U_I - U_K) < (1 - \beta_{AB})(V_I - V_K)$, then $z_K = (1 - \beta_{AB})(V_I - V_K)$. Since $V_K > V_B$, then $(1 - \beta_{AB})(V_I - V_K) < (1 - \beta_{AB})(V_I - V_B) \leq z_B$. Hence, $z_K < z_B$ holds for any value of $\beta_{AB}$ in $[0, 1)$.

- If $\beta_{AB}(U_I - U_K) \geq (1 - \beta_{AB})(V_I - V_K)$, then $z_K = \beta_{AB}(U_I - U_K)$. For $U_K < U_B$ and $V_K > V_B$, we have

$$\beta_{AB}(U_I - U_K) > \beta_{AB}(U_I - U_B),$$
$$(1 - \beta_{AB})(V_I - V_K) < (1 - \beta_{AB})(V_I - V_B).$$

Since $z_K = \beta_{AB}(U_I - U_K) > \beta_{AB}(U_I - U_B)$ and $z_B = \max\{\beta_{AB}(U_I - U_B), (1-\beta_{AB})(V_I - V_B)\}$, then $z_K < z_B$ implies that $\beta_{AB}(U_I - U_K) < (1-\beta_{AB})(V_I - V_B)$. But this should hold for any $U_K$ satisfying $U_K > U_A$, and so, we must have $\beta_{AB}(U_I - U_A) \leq (1-\beta_{AB})(V_I - V_B)$, i.e., $\beta_{AB}(U_I - U_A + V_I - V_B) \leq (V_I - V_B)$. Thus,

$$\beta_{AB} \leq \frac{V_I - V_B}{U_I - U_A + V_I - V_B}. \tag{6.3.9}$$

Thus, if $z_K < z_B$ always holds, the necessary condition for $\beta_{AB}$ is given by (6.3.9).

From (6.3.8) and (6.3.9), we conclude that if both $z_K < z_A$ and $z_K < z_B$ always hold, we have

$$\beta_{AB} = \frac{V_I - V_B}{U_I - U_A + V_I - V_B}.$$

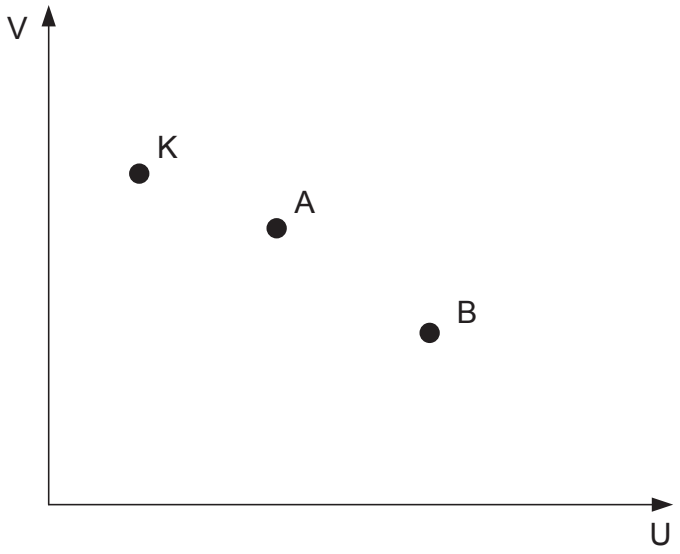
□

Lemma 6.2 provides the necessary condition for $\beta_{AB}$ for generating any new Pareto-optimal point lying strictly between $A$ and $B$ in a simple scenario. However, for any two known Pareto-optimal points $A$ and $B$, we find that $\beta_{AB}$ in (6.3.7) is always a sufficient condition to generate a new Pareto-optimal point between $A$ and $B$ in general when it exists. The following lemma shows that for $\beta$ given by (6.3.7), the corresponding Pareto-optimal point, denoted as $K$ (found by solving BIOPT-L and applying Algorithm 6.1) will be a new Pareto-optimal point between $A$ and $B$ whenever there exists such a Pareto-optimal point between $A$ and $B$, and that $K$ will coincide with either $A$ or $B$ if there does not exist another Pareto-optimal point between $A$ and $B$.

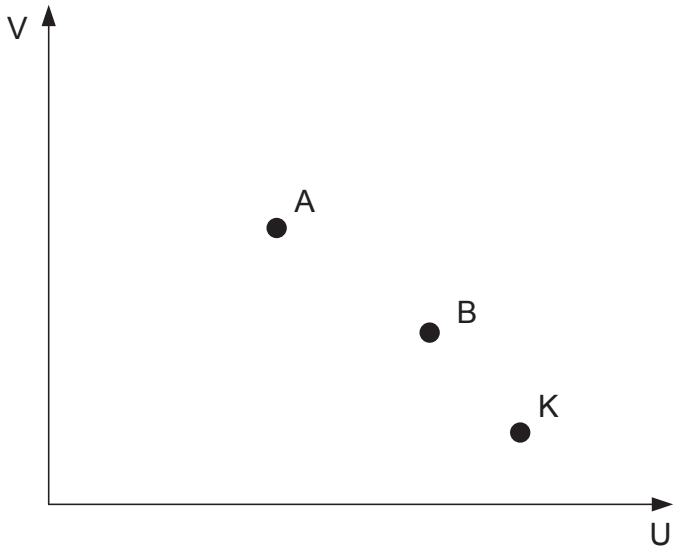
**Lemma 6.3.** If there exists a Pareto-optimal point between $A$ and $B$, then $K$ (corresponding to $\beta_{AB}$ in (6.3.7)) falls strictly between $A$ and $B$; otherwise, $K$ coincides with either $A$ or $B$.

*Proof.* Our proof consists of three steps. We first show that $K$ falls between $A$ and $B$. Next, we show that if there exists other Pareto-optimal points between $A$ and $B$, then $K$ falls strictly between $A$ and $B$. Finally, we show that if no other Pareto-optimal points exist between $A$ and $B$, then $K$ coincides with either $A$ or $B$.

**Step 1.** We first show that $K$ falls between $A$ and $B$. Our proof is based on the contradiction by assuming: (i) $K$ falls to the left of $A$; (ii) $K$ falls to the right of $B$.




(a) The point $K$ falls to the left of $A$.




(b) The point $K$ falls to the right of $B$.

Figure 6.2: Assuming $K$ does not fall between $A$ and $B$.

- (i) In this case, we assume $K$ falls to the left of $A$ (see Fig. 6.2(a)). For any Pareto-optimal point $(U^\dagger, V^\dagger)$ that is between $A$ and $B$, we have $U_K < U^\dagger$ and $V_K > V^\dagger$. The objective value $z_K$ with respect to $(U_K, V_K)$ has the following constraints:

$$z_K \geq \beta_{AB}(U_I - U_K), \quad z_K \geq (1 - \beta_{AB})(V_I - V_K)$$

with at least one constraint satisfying the equivalent condition. For point $(U^\dagger, V^\dagger)$, its objective value $z^\dagger$ has the following constraints:

$$z^\dagger \geq \beta_{AB}(U_I - U^\dagger), \quad z^\dagger \geq (1 - \beta_{AB})(V_I - V^\dagger)$$

with at least one constraint satisfying the equivalent condition.

Since $U_K < U^\dagger$ and $V_K > V^\dagger$, then we have $\beta_{AB}(U_I - U_K) > \beta_{AB}(U_I - U^\dagger)$ and $(1 - \beta_{AB})(V_I - V_K) < (1 - \beta_{AB})(V_I - V^\dagger)$. In order to find the relationships between $z_K$ and $z^\dagger$, we discuss two cases:

  - If $\beta_{AB}(U_I - U^\dagger) \geq (1 - \beta_{AB})(V_I - V^\dagger)$, then $z^\dagger = \beta_{AB}(U_I - U^\dagger)$. Since $z_K \geq \beta_{AB}(U_I - U_K) > \beta_{AB}(U_I - U^\dagger)$, then $z_K > z^\dagger$.
  - If $\beta_{AB}(U_I - U^\dagger) < (1 - \beta_{AB})(V_I - V^\dagger)$, then $z^\dagger = (1 - \beta_{AB})(V_I - V^\dagger)$. Since $\beta_{AB}(U_I - U_K) > \beta_{AB}(U_I - U^\dagger)$ is known, we compare $\beta_{AB}(U_I - U_K)$ with $(1 - \beta_{AB})(V_I - V^\dagger)$. Since $U_K < U_A$, then

$$\begin{aligned} \beta_{AB}(U_I - U_K) &> \beta_{AB}(U_I - U_A) \\ &= \frac{V_I - V_B}{U_I - U_A + V_I - V_B}(U_I - U_A). \end{aligned}$$

    Since $V^\dagger > V_B$, then

$$\begin{aligned} (1 - \beta_{AB})(V_I - V^\dagger) &< (1 - \beta_{AB})(V_I - V_B) \\ &= \frac{U_I - U_A}{U_I - U_A + V_I - V_B}(V_I - V_B). \end{aligned}$$

    Therefore, $z_K \geq \beta_{AB}(U_I - U_K) > (1 - \beta_{AB})(V_I - V^\dagger) = z^\dagger$.

Based on the above discussion, we find that $z_K > z^\dagger$, which means that any Pareto-optimal point between $A$ and $B$ can achieve a smaller $z$ than $K$. This contradicts that $K$ can achieve the minimum $z$ (minimum Chebyshev norm) for BIOPT-L. Therefore, $K$ can not fall to the left of $A$.

- (ii) The discussion for the case that $K$ cannot fall to the right of $B$ (Fig. 6.2(b)) is similar to case $(i)$, we omit it here to conserve space.

From $(i)$ and $(ii)$, we conclude that the Pareto-optimal point $K$ found by setting $\beta_{AB}$ as in (6.3.7) falls between $A$ and $B$.

**Step 2.** From Step 1, we showed that $K$ falls between $A$ and $B$. Here, we show that if there exists new Pareto-optimal points between $A$ and $B$, then the Pareto-optimal point $K$ found by $\beta_{AB}$ will be a new point different from $A$ and $B$. To show this, we only need to show that all Pareto-optimal points that fall strictly between $A$ and $B$ can achieve a smaller $z$ than $A$ and $B$.

For any Pareto-optimal point with $(U^\dagger, V^\dagger)$ that falls strictly between $A$ and $B$, we have $U_A < U^\dagger < U_B$ and $V_A > V^\dagger > V_B$. We define $z^\dagger, z_A, z_B$ as the objective values for BIOPT-L corresponding to Pareto-optimal points $(U^\dagger, V^\dagger)$, $A$ and $B$, respectively. Therefore, we have:

$$
\begin{aligned}
z^\dagger \geq \beta_{AB}(U_I - U^\dagger), \quad z^\dagger \geq (1-\beta_{AB})(V_I - V^\dagger);\\
z_A \geq \beta_{AB}(U_I - U_A), \quad z_A \geq (1-\beta_{AB})(V_I - V_A);\\
z_B \geq \beta_{AB}(U_I - U_B), \quad z_B \geq (1-\beta_{AB})(V_I - V_B).
\end{aligned}
$$

We now show that $z^\dagger < z_A$. We consider the different cases for the relationships between $\beta_{AB}(U_I - U^\dagger)$ and $(1-\beta_{AB})(V^* - V^\dagger)$:

- If $\beta_{AB}(U_I - U^\dagger) \geq (1-\beta_{AB})(V_I - V^\dagger)$, then $z^\dagger = \beta_{AB}(U_I - U^\dagger)$. Since $U_A < U^\dagger$, we have $\beta_{AB}(U_I - U_A) > \beta_{AB}(U_I - U^\dagger)$. Therefore, $z_A \geq \beta_{AB}(U_I - U_A) > \beta_{AB}(U_I - U^\dagger) = z^\dagger$.
- If $\beta_{AB}(U_I - U^\dagger) < (1-\beta_{AB})(V_I - V^\dagger)$: then $z^\dagger = (1-\beta_{AB})(V_I - V^\dagger)$. Since $U_A < U^\dagger$ , we have $\beta_{AB}(U_I - U_A) > \beta_{AB}(U_I - U^\dagger)$. Now we compare $\beta_{AB}(U_I - U_A)$ with

$(1-\beta_{AB})(V_I - V^\dagger)$. Since $V^\dagger > V_B$, then $(1-\beta_{AB})(V_I - V^\dagger) < (1-\beta_{AB})(V_I - V_B) = \frac{(U_I - U_A)}{(U_I - U_A + V_I - V_B)}(V_I - V_B) = \beta_{AB}(U_I - U_A)$. Therefore, $z^\dagger < z_A$.

The proof for $z^\dagger < z_B$ is similar, and we omit it here.

From the above discussion, we find that any Pareto-optimal point $(U^\dagger, V^\dagger)$ that falls strictly between $A$ and $B$ can achieve a smaller $z$ than $A$ and $B$. Therefore, the new Pareto-optimal point $K$ (corresponding to $\beta_{AB}$ in (6.3.7)) will fall strictly between $A$ and $B$.

**Step 3.** We show that if there does not exist any new Pareto-optimal point between Pareto-optimal points $A$ and $B$, then the Pareto-optimal point $K$ will coincides with either $A$ or $B$. In Section 6.3.3, we showed that we can find a Pareto-optimal point for $\beta_{AB}$. From the above, we have shown that this Pareto-optimal point falls in the interval of $A$ and $B$. If there is no other Pareto-optimal point between $A$ and $B$, then the Pareto-optimal point $K$ found by $\beta_{AB}$ can only be either $A$ or $B$. □

The significance of Lemma 6.3 is that it allows us to find new Pareto-optimal points iteratively based on two known Pareto-optimal points. So we can start from two known Pareto-optimal points $\{Q_1, Q_2\}$. Based on these two points, we calculate $\beta$ as in (6.3.7) to find new Pareto-optimal point $Q_3$. We now have two intervals: $\{Q_1, Q_3\}$ and $\{Q_3, Q_2\}$. For each interval, we find its $\beta$ and a new Pareto-optimal point. In the case when the Pareto-optimal point coincides with any of the two end points, we declare that there does not exist a new Pareto-optimal point in this interval. The process continues as long as we can find new Pareto-optimal point for some interval. When the total number of Pareto-optimal points is not large, our algorithm will terminate with all Pareto-optimal points. But when the number of Pareto-optimal points is very large (possibly infinite number of points), we need a way to terminate the iterations. In our algorithm, we set the following termination condition. For the interval between $A$ and $B$, if

$$\max\{U_B - U_A, V_A - V_B\} \leq \varepsilon, \tag{6.3.10}$$

then we stop to find any new Pareto-optimal point in this interval. In the next section, we show that such a termination condition can guarantee a maximum throughput curve that is $\varepsilon$-approximate to

the optimal.

For the $\beta$ values defined by (6.3.7), we have the following result, which is a point of related interest with respect to the stability of solutions to BIOPT-L.

**Property 6.1.** *Suppose* $(U_1, V_1)$, $(U_2, V_2), \cdots, (U_M, V_M)$ *are all Pareto-optimal points between* $Q_1$ *and* $Q_M$ *with* $U_1 < U_2 < \cdots < U_M$*, where* $M$ *could go to infinite. Then, we have the following relationships for the corresponding* $\beta_{(K-1)K}$*:*

$$\beta_{12} < \beta_{23} < \cdots < \beta_{(M-1)M}.$$

*Moreover, for any* $\beta \in (\beta_{(K-1)K}, \beta_{K(K+1)})$*, we have that* $(U_K, V_K)$ *is the corresponding Pareto-optimal point found via the BIOPT-L (note that by Lemma 6.3, for* $\beta = \beta_{(K-1)K}$*, we have that either* $(U_{K-1}, V_{K-1})$ *or* $(U_K, V_K)$ *is optimal, for each* $K = 2, \cdots, M$*).*

*Proof.* **(1).** We first prove that $\beta_{12} < \beta_{23} < \cdots < \beta_{(M-1)M}$.

Based on (6.3.7), we know $\beta_{(K-1)K} = \frac{V_I - V_K}{U_I - U_{K-1} + V_I - V_K}$ and $\beta_{K(K+1)} = \frac{V_I - V_{K+1}}{U_I - U_K + V_I - V_{K+1}}$. Then, they can also be expressed as follows:

$$\beta_{(K-1)K} = \frac{1}{\frac{U_I - U_{K-1}}{V_I - V_K} + 1} \quad \text{and} \quad \beta_{K(K+1)} = \frac{1}{\frac{U_I - U_K}{V_I - V_{K+1}} + 1}.$$

Since $U_{K-1} < U_i < U_{K+1}$ and $V_{K-1} > V_K > V_{K+1}$, then $\frac{U_I - U_{K-1}}{V_I - V_K} > \frac{U_I - U_K}{V_I - V_{K+1}}$, and so we can conclude that $\beta_{(K-1)K} = \frac{1}{\frac{U_I - U_{K-1}}{V_I - V_K} + 1} < \frac{1}{\frac{U_I - U_K}{V_I - V_{K+1}} + 1} = \beta_{K(K+1)}$.

Since $\beta_{(K-1)K} < \beta_{K(K+1)}$ for any adjacent Pareto-optimal points, we have

$$\beta_{12} < \beta_{23} < \cdots < \beta_{(M-1)M}.$$

**(2).** We next prove that for any $\beta$ where $\beta_{(K-1)K} < \beta < \beta_{K(K+1)}$, the corresponding Pareto-optimal point is $(U_K, V_K)$, i.e., $(U_K, V_K)$ achieves the minimum $z$ value. For $(U_K, V_K)$, the objective value for BIOPT-L is $z_K$. For any other Pareto-optimal point $(U_R, V_R)$, we define its objective value as $z_R$. For $z_K$, we have

$$z_K \geq \beta(U_I - U_K), \quad z_K \geq (1 - \beta)(V_I - V_K).$$

If $\beta(U_I - U_K) \geq (1-\beta)(V_I - V_K)$, then $z_K = \beta(U_I - U_K)$. For $(U_R, V_R)$, we consider two cases:

**(i)** If $U_R < U_K$ and $V_R > V_K$, then $z_K = \beta(U_I - U_K) < \beta(U_I - U_R) \leq z_R$.

**(ii)** If $U_R > U_K$ and $V_R < V_K$, since $\beta < \beta_{K(K+1)}$, then we have

$$z_K = \beta(U_I - U_K) < \beta_{K(K+1)}(U_I - U_K) = \frac{(V_I - V_{K+1})(U_I - U_K)}{U_I - U_K + V_I - V_{K+1}},$$

$$z_R = (1-\beta)(V_I - V_R) > (1-\beta_{K(K+1)})(V_I - V_R) = \frac{(U_I - U_K)(V_I - V_{K+1})}{U_I - U_K + V_I - V_{K+1}}.$$

Therefore, $z_K < z_R$.

If $\beta(U_I - U_K) \leq (1-\beta)(V_I - V_K)$, then $z_K = (1-\beta)(V_I - V_K)$.

We also consider two cases for $(U_R, V_R)$:

**(i)** $U_R > U_K$ and $V_R < V_K$ : Thus $z_K = (1-\beta)(V_I - V_K) < (1-\beta)(V_I - V_R) \leq z_R$.

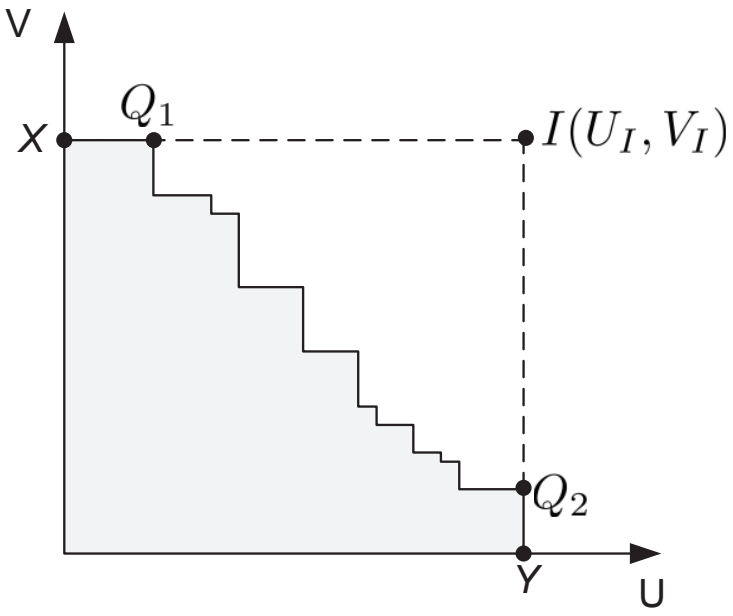
**(ii)** $U_R < U_K$ and $V_R > V_K$ : Since $\beta > \beta_{(K-1)K}$, then we have:

$$z_K = (1-\beta)(V_R - V_K) < (1-\beta_{(K-1)K})(V_I - V_K) = \frac{(U_I - U_{K-1})(V_I - V_K)}{U_I - U_{K-1} + V_I - V_K},$$

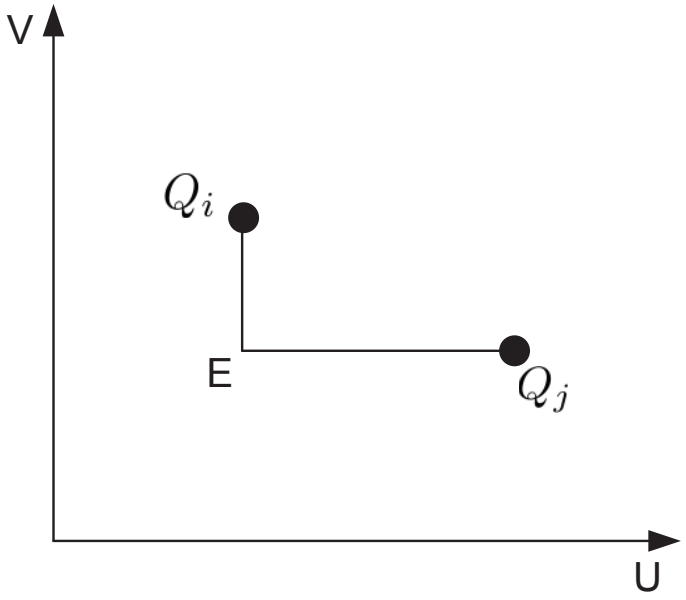
$$z_R \geq \beta(U_I - U_R) > \beta_{(K-1)K}(U_I - U_R) = \frac{(V_I - V_K)(U_I - U_R)}{U_I - U_{K-1} + V_I - V_K}$$

Therefore, $z_K < z_R$, which is also a contradiction that $z_R$ is the optimal solution.

Hence, for any Pareto-optimal point $R$, we have $z_K < z_R$, which implies that $(U_K, V_K)$ is the optimal solution of BIOPT-L for $\beta \in (\beta_{(K-1)K}, \beta_{K(K+1)})$. □




(a) The ideal point and two starting Pareto-optimal points.




(b) Throughput curve between any two consecutive $Q_i$ and $Q_j$.

Figure 6.3: Assuming $K$ does not fall between $A$ and $B$.

### 6.3.5 Main Result

We are now ready to describe the complete algorithm to find the necessary number of Pareto-optimal points that can be used to construct a throughput curve. As shown in Fig. 6.3(a), we start with our ideal point $I(U_I, V_I)$ and weakly Pareto-optimal points $X$ and $Y$ with $(0, V_I)$ and $(U_I, 0)$, respectively. By using Algorithm 6.1, we can find the Pareto-optimal points $Q_1$ and $Q_2$ corresponding to $X$ and $Y$, respectively. Note that when $X$ and $Y$ are already Pareto-optimal, then $Q_1$ and $Q_2$ will coincide with $X$ and $Y$, respectively. Starting from the interval with two end points $Q_1$ and $Q_2$, we can find other new Pareto-optimal points iteratively. The iteration terminates when there is no new Pareto-optimal point for each interval or the interval is sufficiently small (as in (6.3.10)). Since there is a non-zero continuous interval between any two neighboring Pareto-optimal points, the total number of Pareto-optimal points in $\mathcal{G}$ is thus finite. Based on the weakly Pareto-optimal points $X$ and $Y$, and Pareto-optimal points $Q_1, \cdots, Q_2$ that we have found in the iterations, we have a throughput curve as follows: i) connect $X$ and $Q_1$ with a line, ii) make an "L"-shape connection between any two consecutive Pareto-optimal points between $Q_i$ and $Q_j$ as shown in Fig. 6.3(b), and iii) connect $Q_2$ and $Y$ with a line. Fig. 6.4 summarizes our discussions.

**Theorem 3.** The throughput curve from Figure 6.4 approximates the optimal bicriteria throughput curve with the approximation error no more than $\varepsilon$.

*Proof.* We consider any two adjacent points $Q_i$ and $Q_j$ (see Fig. 6.5).

- If there does not exist any other Pareto-optimal points between $Q_i$ and $Q_j$, then the curve $Q_i - E - Q_j$ is exact the optimal throughput curve, since all points on this curve are weakly Pareto-optimal points.
- If there exist Pareto-optimal points between $Q_i$ and $Q_j$, for any one of these Pareto-optimal points, say $R$ with $(U_R, V_R)$, we have $U_{Q_i} < U_R < U_{Q_j}$ and $V_{Q_j} < V_R < V_{Q_i}$. When we use $D_1$ with $(U_R, V_{Q_j})$ (or $D_2$ with $(U_{Q_i}, V_R)$) to approximate $R$, then we have $\max\{U_R - U_{Q_i}, V_R - V_{Q_j}\} < \max\{U_{Q_j} - U_{Q_i}, V_{Q_i} - V_{Q_j}\} \leq \varepsilon$. Then, the approximation error by using $D_1$ (or $D_2$) to approximate $R$ will be no more than $\varepsilon$.

**Plotting Throughput Curve**

1. **Initialization:** Find ideal point $I$, weakly Pareto-optimal points $X$ and $Y$, and corresponding Pareto-optimal points $Q_1$ and $Q_2$.
2. Set $\mathcal{Z} = \{\{Q_1, Q_2\}\}$, and $\mathcal{G} = \{Q_1, Q_2\}$.
3. while ($\mathcal{Z} \neq \emptyset$) {
4. Take an interval, say $\{Q_i, Q_j\}$, from set $\mathcal{Z}$.
5. If $(\max\{U_{Q_j} - U_{Q_i}, V_{Q_i} - V_{Q_j}\} > \varepsilon)$ {
6. Compute $\beta_{Q_iQ_j}$ based on (6.3.7) to find new Pareto-optimal point $Q_k$.
7. If $Q_k$ coincides with $Q_i$ or $Q_j$, then $Q_i$ and $Q_j$ are two adjacent Pareto-optimal points.
8. Otherwise, $Q_k$ is a new Pareto-optimal point. $\mathcal{Z} = \mathcal{Z} \cup \{\{Q_i, Q_k\}, \{Q_k, Q_j\}\}$, $\mathcal{G} = \mathcal{G} \cup \{Q_k\}$.}
9. Remove $\{Q_i, Q_j\}$ from $\mathcal{Z}$. }
10. Draw throughput curve based on $X$, $Y$, and all Pareto-optimal points in $\mathcal{G}$.

Figure 6.4: Pseudo-code of an approximation algorithm to find $(1 - \varepsilon)$-optimal throughput curve.

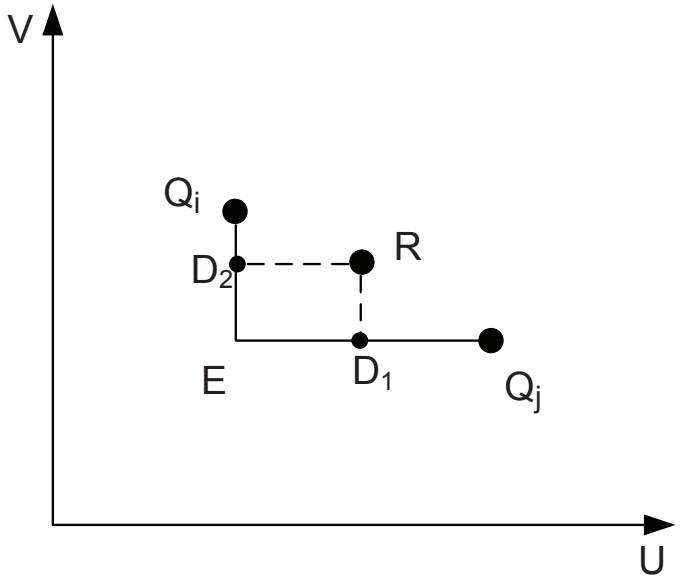



Figure 6.5: The Pareto-optimal point $R$ is represented by $D_1$ (or $D_2$) with $\varepsilon$-approximation.

Therefore, when we use $Q_i - E - Q_j$ to approximate the Pareto-optimal curve between $Q_i$ and $Q_j$, the approximation error will be no more than $\varepsilon$. □

## 6.4 A Case Study

In this section, we perform a numerical study on a primary and secondary networks. Our goal is twofold. First, we want to demonstrate how our algorithm finds throughput curve for the bicriteria optimization problem. Second, we want to compare the throughput region (the area under the throughput curve) under the UPS policy to that under the interweave.

### 6.4.1 Simulation Setting

We consider a randomly generated 15-node primary network and 15-node secondary network in a $100 \times 100$ area. For generality, we normalize the units for distance, bandwidth, power and data rate with appropriate dimensions. The location of each node is shown in Fig. 6.6. We assume that

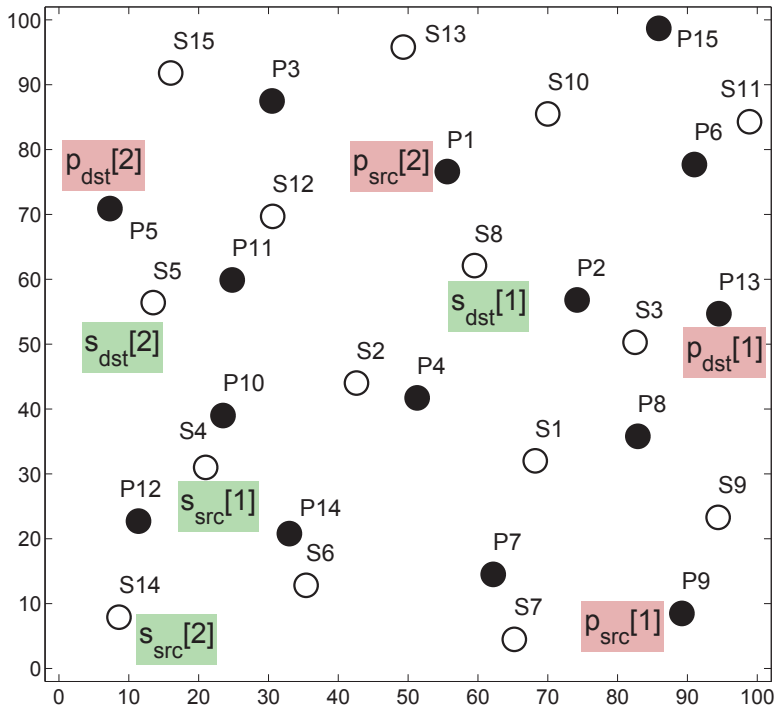



Figure 6.6: The locations of a 15-node primary network and a 15-node secondary network.

there are two primary sessions in the primary network and two secondary sessions in the secondary network. The source and destination nodes for each session are randomly chosen in their networks and are also shown in Fig. 6.6.

We assume there are two channels in the primary network ($\mathcal{B} = \{1, 2\}$), with the bandwidth of each channel being $W = 10$. A time frame is divided into four time slots ($T = 4$). The transmission power spectral density $\rho_i$ for each node $i \in \mathcal{N}$ is 1, the path loss index $\gamma$ is 4, the antenna related constant $\lambda$ is 1, and the ambient Gaussian noise density $N_0 = 10^{-6}$. We assume a node' transmission range and interference range are 30 and 50, respectively, in both primary and secondary networks. We set the approximation error for objective $\varepsilon = 0.1$.

### 6.4.2 Throughput Curve

For the above network setting, we apply our algorithm to BIOPT-L and find a sequence of Pareto-optimal points. We first set the ideal point $I$ to $(15.2235, 10.4497)$. Then we find two starting

Table 6.2: New Pareto-optimal point that is found by two known Pareto-optimal points in each iteration. “PO” represents Pareto-optimal points.

| Iteration | $Q_i$ | $Q_j$ | $\beta_{Q_iQ_j}$ | New PO point |
|---|---|---|---|---|
| 1 | (0, 10.44) | (15.22, 0) | 0.407 | (10.5, 6.96) |
| 2 | (0, 10.44) | (10.5, 6.96) | 0.186 | (5.025, 8.11) |
| 3 | (10.501, 6.96) | (15.223, 0) | 0.688 | (12.28, 4.53) |
| 4 | (0, 10.44) | (5.02, 8.11) | 0.133 | (4.2, 9.07) |
| 5 | (5.025, 8.11) | (10.5, 6.96) | 0.254 | (7.01, 8.0) |
| 6 | (10.5, 6.96) | (12.289, 4.53) | 0.556 | (11.8, 6.00) |
| 7 | (12.28, 4.53) | (15.22, 0) | 0.780 | (13.16, 3.55) |
| 8 | (0, 10.44) | (4.206, 9.072) | 0.082 | (3.477, 9.26) |
| 9 | (4.20, 9.07) | (5.025, 8.112) | 0.175 | (4.88, 8.25) |
| 10 | (5.02, 8.11) | (7.01, 8.0) | 0.193 | (5.13, 8.02) |
| 11 | (7.01, 8.0) | (10.50, 6.96) | 0.298 | (8.46, 7.46) |
| 12 | (10.5, 6.96) | (11.8, 6.00) | 0.485 | NO |
| 13 | (11.8, 6.00) | (12.28, 4.53) | 0.572 | NO |
| 14 | (12.28, 4.53) | (13.16, 3.55) | 0.702 | NO |
| 15 | (13.164, 3.554) | (15.223, 0) | 0.835 | NO |
| 16 | (0, 10.44) | (3.47, 9.26) | 0.072 | (0.93, 9.33) |
| 17 | (3.47, 9.26) | (4.2, 9.07) | 0.006 | NO |
| 18 | (4.2, 9.07) | (4.88, 8.25) | 0.166 | (4.76, 8.36) |
| 19 | (4.88, 8.25) | (5.02, 8.11) | 0.184 | (4.99, 8.13) |
| 20 | (5.02, 8.11) | (5.13, 8.02) | 0.191 | (5.16, 8.0 ) |
| 21 | (5.13, 8.02) | (7.01, 8.0) | 0.195 | (5.16, 8.0) |
| 22 | (7.01, 8.0) | (8.46, 7.46) | 0.191 | NO |
| 23 | (8.46, 7.46) | (10.5, 6.96) | 0.340 | (9.06, 7.27) |
| 24 | (0, 10.44) | (0.93, 9.33) | 0.068 | (0.87, 9.4) |
| 25 | (0.93, 9.33) | (3.47, 9.26) | 0.076 | (1.0, 9.27) |
| 26 | (4.2, 9.07) | (4.76, 8.36) | 0.158 | (4.68, 8.45) |
| 27 | (4.76, 8.36) | (4.88, 8.25) | 0.173 | (4.86, 8.27) |
| 28 | (4.88, 8.25) | (4.99, 8.13) | 0.182 | (4.97, 8.15) |
| 29 | (4.99, 8.13) | (5.02, 8.11) | — | — |
| 30 | (5.13, 8.02) | (5.16, 8.0) | — | — |
| 31 | (5.16, 8.0) | (7.01, 8.0) | 0.135 | NO |
| 32 | (8.46, 7.46) | (9.06, 7.27) | 0.319 | (8.88, 7.45) |

| Iteration | $Q_i$ | $Q_j$ | $\beta_{Q_iQ_j}$ | New PO point |
|---|---|---|---|---|
| 33 | (9.06, 7.27) | (10.5, 6.96) | 0.361 | (9.26, 7.07) |
| 34 | (0, 10.44) | (0.87, 9.4) | 0.064 | (0.81, 9.45) |
| 35 | (0.87, 9.4) | (0.93, 9.33) | — | — |
| 36 | (0.93, 9.33) | (1.0, 9.27) | — | — |
| 37 | (1.0, 9.27) | (3.47, 9.26) | 0.005 | NO |
| 38 | (4.20, 9.07) | (4.68, 8.45) | 0.153 | (4.6, 8.52) |
| 39 | (4.68, 8.45) | (4.76, 8.36) | 0.054 | NO |
| 40 | (8.461, 7.461) | (8.881, 7.450) | 0.211 | NO |
| 41 | (8.88, 7.45) | (9.06, 7.27) | 0.333 | (9.0, 7.33) |
| 42 | (9.06, 7.27) | (9.26, 7.07) | 0.354 | (9.19, 7.14) |
| 43 | (9.26, 7.07) | (10.5, 6.96) | 0.369 | (9.33, 7.0) |
| 44 | (0, 10.44) | (0.81, 9.45) | 0.061 | (0.76, 9.5) |
| 45 | (0.81, 9.45) | (0.87, 9.4) | 0.044 | NO |
| 46 | (4.2, 9.07) | (4.6, 8.52) | 0.148 | (4.54, 8.58) |
| 47 | (4.6, 8.52) | (4.68, 8.45) | — | — |
| 48 | (9.26, 7.07) | (9.33, 7.0) | — | — |
| 49 | (9.33, 7.0) | (10.5, 6.96) | 0.372 | (9.36, 6.97) |
| 50 | (0, 10.44) | (0.76, 9.5) | 0.058 | (0.72, 9.55) |
| 51 | (0.76, 9.5) | (0.8, 9.45) | — | — |
| 52 | (4.2, 9.07) | (4.54, 8.58) | 0.144 | (4.49, 8.63) |
| 53 | (4.54, 8.58) | (4.6, 8.52) | — | — |
| 54 | (9.33, 7.0) | (9.36, 6.97) | — | — |
| 55 | (9.36, 6.97) | (10.5, 6.96) | 0.272 | NO |
| 56 | (0, 10.44) | (0.72, 9.55) | 0.055 | (0.52, 9.58) |
| 57 | (0.72, 9.55) | (0.76, 9.5) | — | — |
| 58 | (4.2, 9.07) | (4.49, 8.63) | 0.141 | (4.45, 8.67) |
| 59 | (4.49, 8.63) | (4.54, 8.58) | — | — |
| 60 | (0, 10.44) | (0.52, 9.58) | 0.036 | NO |
| 61 | (0.52, 9.58) | (0.72, 9.55) | 0.037 | NO |
| 62 | (4.2, 9.07) | (4.45, 8.67) | 0.138 | (4.42, 8.71) |
| 63 | (4.45, 8.67) | (4.49, 8.63) | — | — |
| 64 | (4.2, 9.07) | (4.42, 8.71) | 0.136 | (4.39, 8.74) |
| 65 | (4.42, 8.71) | (4.45, 8.67) | — | — |
| 66 | (4.2, 9.07) | (4.39, 8.74) | 0.086 | NO |
| 67 | (4.39, 8.74) | (4.42, 8.71) | — | — |

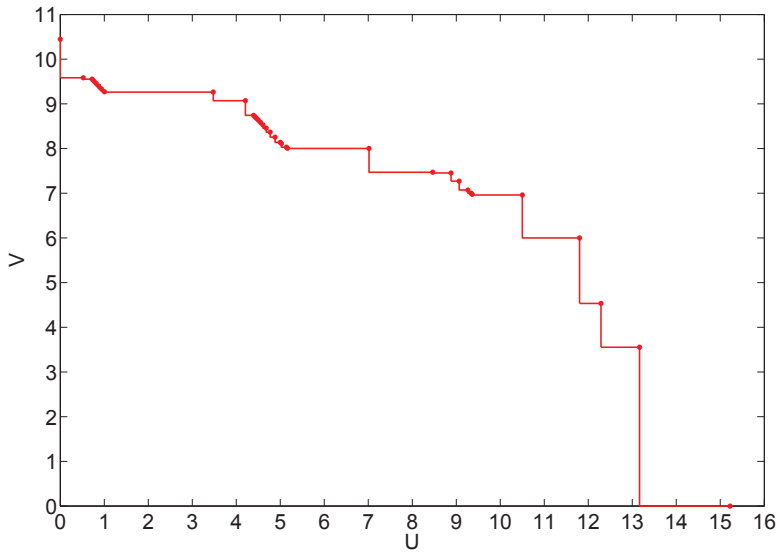



Figure 6.7: The throughput curve found by our algorithm.

Pareto-optimal points $Q_1$ and $Q_2$ as $(0, 10.4497)$ and $(15.2235, 0)$, respectively. Based on these two starting Pareto-optimal points, Table 6.2 shows the results from our iterations. For each iteration, we have two Pareto-optimal points $Q_i$ and $Q_j$. Based on these two points, we find $\beta_{Q_iQ_j}$ and the corresponding new Pareto-optimal point. In iteration 12–15, 17, 22, 31, 37, 39, 40, 45, 55, 60, 61, and 66, there does not exist a new Pareto-optimal point for the corresponding interval. In iteration 29, 30, 35, 36, 47, 48, 51, 53, 54, 57, 59, 63, 65, and 67, we stop finding Pareto-optimal points since the corresponding intervals between $Q_i$ and $Q_j$ are smaller than $\varepsilon$.

Based on the Pareto-optimal points in Table 6.2, we plot the throughput curve as shown in Fig. 6.7. On this curve, the intervals corresponding to iterations 29, 30, 35, 36, 47, 48, 51, 53, 54, 57, 59, 63, 65, and 67 have $\varepsilon$-approximation to the optimal. For intervals corresponding to iterations 12–15, 17, 22, 31, 37, 39, 40, 45, 55, 60, 61 and 66, the throughput curve is optimal. These can be validated by choosing any point $(U^*, V^*)$ on this curve and compare it with the corresponding optimal point. This optimal point can be obtained by solving BIOPT by setting the primary network throughput $U = U^*$. Then, we compare the maximum secondary network throughput value $V$ with that we found from the curve. We first validate the points that locate within the intervals that achieve the $\varepsilon$-approximation. We set the primary network throughput

$U^* = 4.65$, and solve BIOPT. The maximum secondary network throughput is $V = 8.476$. Based on the curve in Fig. 6.7, we can find the secondary network throughput $V^* = 8.457$. The difference between $V = 8.476$ and $V^* = 8.457$ is 0.019, which is smaller than $\varepsilon = 0.1$. For any point located in the intervals corresponding to iterations 29, 30, 35, 36, 47, 48, 51, 53, 54, 57, 59, 63, 65, and 67, we can obtain similar results. Next, we validate the results that are located in the intervals corresponding to iterations 12–15, 17, 22, 31, 37, 39, 40, 45, 55, 60, 61 and 66. We choose the primary network throughput $U^* = 2.0$, and solve BIOPT by setting $U = U^*$. The obtained maximum secondary network throughput is $V = 9.270$. Based on the curve in Fig. 6.7, we can find the maximum secondary network throughput $V^* = 9.270$, which is the same as the optimal and the difference is smaller than $\varepsilon$. We can repeat the validation for any point located in intervals corresponding to iterations 12–15, 17, 22, 31, 37, 39, 40, 45, 55, 60, 61 and 66, and obtain the same conclusion. We omit to show those results here to conserve space.

Not shown in Table 6.2 are the feasible solutions for the Pareto-optimal points. In particular, in iterations 1–11, 16, 18–21, 23–28, 32–34, 38, 41–44, 46, 49, 50, 52, 56, 58, 62, and 64. we have found a Pareto-optimal solution $\alpha^\dagger$ when we solve the new Pareto-optimal point $(U^\dagger, V^\dagger)$. We now show how to find a feasible solution for *any* point on the throughput curve in Fig. 6.7.

Consider a point (10, 6.96) on the throughput curve in Fig. 6.7. This point falls in the interval in iteration 55. Since this point is not a Pareto-optimal point, we do not know its feasible solution $\alpha^* = \{\mathbf{x}^*, \mathbf{f}^*, \hat{\mathbf{f}}^*, \mathbf{r}^*, \hat{\mathbf{r}}^*, \mathbf{r}^*_{\min}, \hat{\mathbf{r}}^*_{\min}\}$. We show how to construct a feasible solution for point (10, 6.96) based on the solution for its corresponding Pareto-optimal point $(10.501, 6.96)$. For $(10.501, 6.96)$, denote its solution as $\alpha^\dagger = \{\mathbf{x}^\dagger, \mathbf{f}^\dagger, \hat{\mathbf{f}}^\dagger, \mathbf{r}^\dagger, \hat{\mathbf{r}}^\dagger, \mathbf{r}^\dagger_{\min}, \hat{\mathbf{r}}^\dagger_{\min}\}$. For $\alpha^*$, we can use the same scheduling as in $\alpha^\dagger$ for both primary and secondary sessions, i.e., $\mathbf{x}^* = \mathbf{x}^\dagger$. For flow routing and data rate, there is no change for the secondary session, i.e., $\mathbf{f}^* = \mathbf{f}^\dagger$ and $\mathbf{r}^* = \mathbf{r}^\dagger$. But for the primary sessions, their throughput need to be adjusted, although their routing topology do not change. Specifically, we adjust the primary session throughput from 10.501 in $\alpha^\dagger$ to 10 in $\alpha^*$, which will affect data rate on each link $\hat{\mathbf{f}}^*$. Therefore, we obtain a feasible solution $\alpha^*$ for point (10, 6.96). For any point on this curve, we can use this method to construct one feasible solution based on the solution of its corresponding Pareto-optimal point.

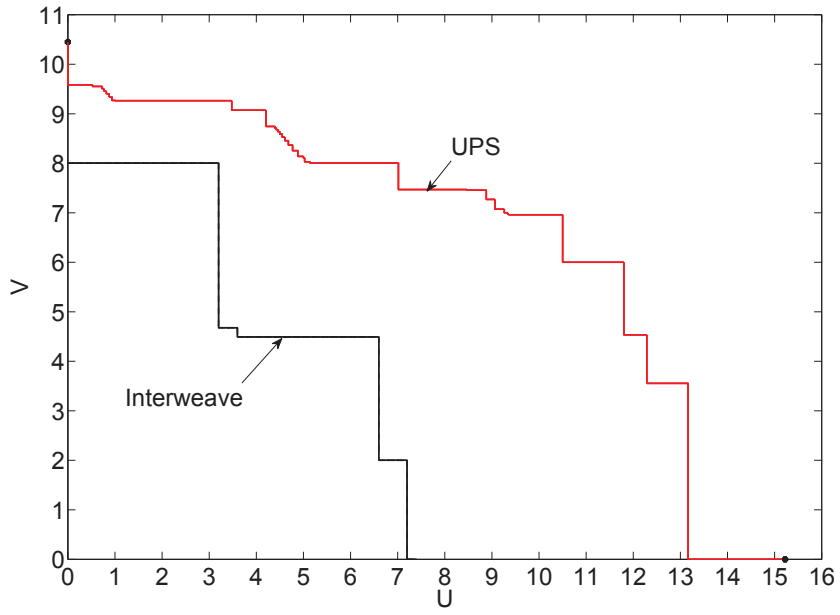



Figure 6.8: A comparison of the throughput region under the UPS policy and the interweave paradigm.

### 6.4.3 Comparison to Other Paradigms

We now compare the UPS's throughput region (the area under the throughput curve) with other paradigms (i.e., underlay and interweave). Since the underlay paradigm requires interference cancellation capabilities (e.g, MIMO [75, 85, 86]) at the physical layer that is beyond what we have assumed for each node for UPS, it is not appropriate (or fair) to make such a comparison. So we will limit our comparison of UPS to the interweave paradigm [22]. Under interweave, the primary nodes use its network and spectrum resource without considerations of the secondary nodes. The secondary nodes are allowed to use a spectrum band only when the primary nodes are not using it. There is no node-level cooperation between the two networks. Fig. 6.8 shows the throughput curves under the UPS and interweave paradigms. The throughput region (in terms of its area size) for the UPS policy is 2.64 times of that for the interweave. We also run 100 instances with different network settings to find the throughput curves between primary and secondary networks with our algorithm. The results are consistent and show that the throughput regions for the UPS policy are always much larger than those under the interweave paradigm.

## 6.5 Chapter Summary

Node-level (data plane) cooperation between the primary and secondary networks adds a new dimension for efficient spectrum sharing. In this chapter, we investigate achievable throughput region when the primary and secondary nodes are allowed to cooperate and forward each other's traffic. The achievable throughput region is characterized by the so-called optimal throughput curve. To find the optimal throughput curve, we formulate a multicriteria optimization problem and developed a novel solution based on weighted Chebyshev norm. Our solution is able to find a sequence of new Pareto-optimal points through iterations. We further show that our throughput curve is an $\varepsilon$-approximation to the optimal. Through a case study, we show that the throughput region under the UPS policy (with node-level cooperation) is substantially larger than that under the interweave paradigm (where there is no node-level cooperation).

# Chapter 7

# Coexistence between Wi-Fi and LTE on Unlicensed Spectrum

## 7.1 Introduction

Today there are over 350 million cellular subscribers in the US and 70% of them possess smartphones. The data traffic carried by these subscribers has exceeded 4.8 exabyte per year and is growing at 50% annually. But the radio frequency spectrum that can be used for wireless communications is a finite and extremely valuable resource. With the proliferation of new wireless applications, the use of the radio spectrum has intensified to the point that new spectrum policies are needed.

On the other hand, there is a significant amount of unlicensed spectrum available. For example, in the 5 GHz band, there is a close to 500 MHz of spectrum bandwidth available (e.g., [5.15, 5.25] GHz and [5.47, 5.85] GHz in the US). Currently, the widely deployed wireless technology on the 5 GHz unlicensed band is Wi-Fi. The idea of deploying cellular over unlicensed spectrum is attractive for telecommunications carriers as it allows them to increase overall capacity without paying billions of dollars that they do for a licensed spectrum. Already, US cellular operators such

as Verizon and T-Mobile are exploring this possibility and making plans to deploy LTE Unlicensed (LTE-U [18, 47, 67]) technology in the unlicensed bands (especially in the 5 GHz band). For the Wi-Fi community, there is a grave concern that the entry of LTE-U (and LAA [40]) protocols will degrade the service quality of Wi-Fi devices since LTE does not employ CSMA (or listen-before-talk (LBT)), which is the key technology for Wi-Fi users to access and share the spectrum. When Wi-Fi and LTE operate in the same unlicensed band, the transmission of Wi-Fi users will be deferred by LTE signals, which leads to degradation to Wi-Fi throughput. In [29, 53, 69], experimental results showed that Wi-Fi throughput may be reduced by 90% when interfered by LTE. This is unfair to Wi-Fi and has led to protest by the Wi-Fi Alliance. To address this issue, the cellular carriers have proposed more friendly coexistence between Wi-Fi and LTE. In Section 7.8, we review related work in this area and point out some fundamental issues with the proposed coexistence schemes.

Instead of taking any side in the coexistence debate, we take a neutral approach to gain a fundamental understanding of coexistence between the two technologies. The novelty of our approach is to focus on user satisfaction rather than following either Wi-Fi or cellular carriers' perspective. This approach is sensible as an important goal of any Wi-Fi or cellular carrier to maximize user satisfaction (besides making a profit). In this chapter, we ask the following two fundamental questions: (1) From user satisfaction perspective, is there any benefit in coexistence between Wi-Fi and LTE beyond just deploying Wi-Fi? (2) If there is a benefit for coexistence, then how to achieve such benefit in practice?

We address the above two questions by studying several deployment and spectrum sharing strategies for Wi-Fi and LTE. We consider a wireless service area on the order of a picocell which can be served by one LTE base station (BS) or multiple Wi-Fi APs (see Figure 7.1). For a user, it has the option to use Wi-Fi for free or LTE for a fee. We introduce a user satisfaction function under Wi-Fi and LTE and study the problem of how to maximize total user satisfaction among all users under different Wi-Fi and LTE deployment scenarios and spectrum sharing policies. Through rigorous mathematical modeling and extensive simulation studies, we find that in terms of maximizing total user satisfaction function, there does not appear to be any benefit with coexistence

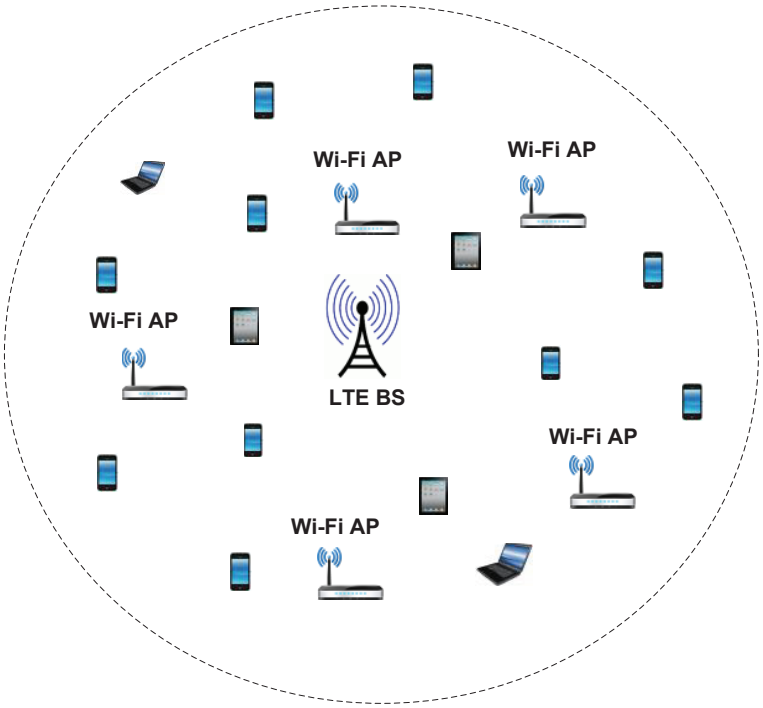



Figure 7.1: The coexistence of Wi-Fi and LTE in a picocell-sized area.

between Wi-Fi and LTE when the unlicensed spectrum is partitioned statically between Wi-Fi and LTE. This is interesting as it suggests that one might just deploy Wi-Fi without LTE in the unlicensed spectrum, when the objective is to maximize total user satisfaction. This finding serves as a powerful counter argument to some of the telecom carriers' proposal to enter the unlicensed spectrum space through static partitioning of the unlicensed band between Wi-Fi and LTE. On the other hand, we find that there is a significant benefit in deploying adaptive spectrum partitioning between Wi-Fi and LTE. That is, the total user satisfaction can be significantly increased when deploying adaptive spectrum partitioning between Wi-Fi and LTE.

Based on the above findings, we conclude that adaptive spectrum partitioning is the only viable approach for coexistence between Wi-Fi and LTE in the unlicensed spectrum. However, such fully adaptive spectrum partitioning is based on global optimization, which means that a user may have to change its service provider whenever there is a new user request arrival or a departure of an existing users. This is not practical as frequent changes of service provider for a user is

disruptive at the application layer. To address this problem, we propose a practical semi-adaptive algorithm to implement fully adaptive spectrum allocation without affecting existing users' service providers. Through performance evaluation, we show the performance of the proposed practical semi-adaptive algorithm is highly competitive when compared to fully adaptive spectrum partition.

The remainder of this chapter is organized as follows. In Section 7.2, we propose a network architecture for coexistence between Wi-Fi and LTE. In Sections 7.3, 7.4, and 7.5, we present three Wi-Fi and LTE service deployment strategies: (1) Wi-Fi only (no LTE); (2) static spectrum partitioning; (3) fully adaptive spectrum partitioning. Section 7.6 presents extensive numerical results to compare the three strategies. In Section 7.7, we propose a practical semi-adaptive algorithm to implement fully adaptive spectrum partitioning and present its performance results. Section 7.8 presents related work and Section 7.9 concludes this chapter.

## 7.2 Network Architecture

In this section, we describe a system architecture for coexistence and spectrum sharing between Wi-Fi and LTE networks. As an example, we consider wireless access at an airport or a similar area on the scale of a picocell. We assume this can be served by one LTE base station (BS) and multiple Wi-Fi APs. As shown in Figure 7.1, the LTE BS has coverage of all users in the area while a Wi-Fi AP can only cover a smaller sub-area (and thus multiple Wi-Fi APs are needed to cover the entire area). Suppose there is a set of users (e.g., laptops, cellphones) in this area wishing to access network service. A user may choose either the LTE BS or one of the Wi-F APs in her neighborhood. If a user chooses LTE, then her subscribed rate will be guaranteed during the lifetime of the connection, but for a price per unit of data rate. On the other hand, if a user chooses Wi-Fi, then her data rate cannot be guaranteed, but the service is free. We assume that each user has her particular financial means (affordability). This affordability is non-negative and reflects how much money a user is willing to pay to access the network. If it is zero, this user will only access the Wi-Fi network; otherwise, it can access either the LTE or the Wi-Fi network.

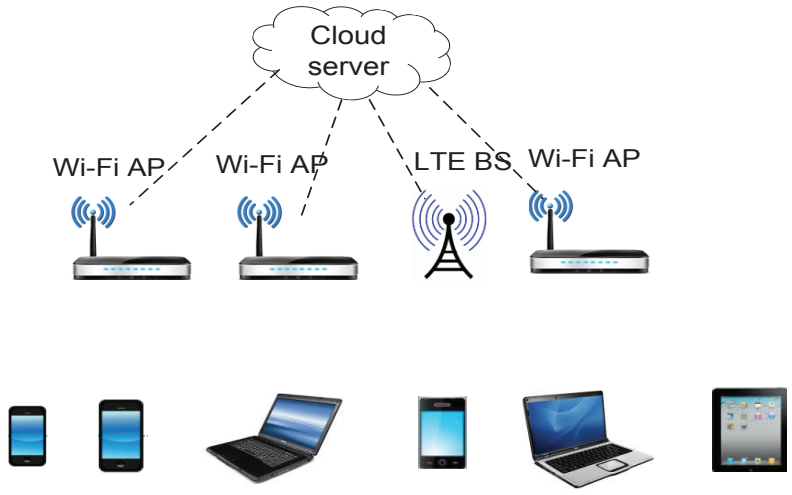



Figure 7.2: A cloud-based control plane that coordinates spectrum sharing between Wi-Fi APs and LTE BS.

Figure 7.2 shows a conceptual of control plane. We assume there is a cloud server deployed at the backend, which connects to both the Wi-Fi APs and LTE BS. The cloud server has powerful computation capability and can compute optimal solutions to maximize users satisfaction based on input from the Wi-Fi and LTE. By default, a user's request for network access goes to a Wi-Fi AP, which will relay the request to the centralized cloud server. Upon receiving the request, the cloud server finds the optimal solution (Wi-Fi AP or LTE service selection and associated spectrum allocation) for the user with the goal of maximizing total users satisfaction. For a user with zero affordability, the cloud server will only assign one of the Wi-Fi APs to her. Otherwise, the cloud server can assign either a Wi-Fi AP or the LTE BS to this user. After making the optimal decision on service selection and spectrum allocation, the cloud server sends the optimal solution to the Wi-Fi APs and LTE BS, which will implement this solution.

In this network architecture, denote $\mathcal{A}$ as the set of Wi-Fi APs and $L$ as the LTE BS. Denote $\mathcal{N}$ as the set of users in this area and denote $\mathcal{N}_i$ as a subset of users that are within the CSMA contention range of user $i$. That is, user $i$ is allowed to transmit only when the set of users in $\mathcal{N}_i$ is not transmitting. Define $\mathcal{A}_i$ as a subset of Wi-Fi APs that covers user $i$. We assume the bandwidth of unlicensed spectrum in the area is $B$. Denote $p$ as the price per unit of data rate imposed by

Table 7.1: Notation

| | |
|---|---|
| $L$ | LTE base station. |
| $\mathcal{N}$ | The set of users in the area. |
| $\mathcal{A}$ | The set of Wi-Fi APs in the area. |
| $\mathcal{N}_i$ | The set of users that are within the CSMA contention range of user $i \in \mathcal{N}$. |
| $\mathcal{A}_i$ | The set of Wi-Fi APs the covers user $i$. |
| $p$ | The price charged by LTE per unit of data rate. |
| $P_i$ | The maximum price for data rate that a user $i$ can afford. |
| $B$ | The total available bandwith for unlicense spectrum. |
| $B_W$ | Bandwidth in unlicense spectrum allocated for Wi-Fi. |
| $B_L$ | Bandwidth in unlicense spectrum allocated for LTE. |
| $B_i^L$ | Bandwidth assigned to user $i$ under LTE. |
| $x_{ij}$ | A binary variable indicating whether or not user $i$ is assigned to Wi-Fi AP $j$. |
| $x_{iL}$ | A binary variable indicating whether or not user $i$ is assigned to LTE BS $L$. |
| $r_{ij}^W$ | The achievable uplink throughput for user $i$ when served by Wi-Fi AP $j$. |
| $r_i^L$ | The achievable uplink throughout for user $i$ when served by LTE. |
| $S_W$ | User's satisfaction coefficient per unit of data rate under Wi-Fi. |
| $S_L$ | User's satisfaction coefficient per unit of data rate under LTE. |
| $\alpha$ | The spectrum efficiency for Wi-Fi. |
| $Q_i^W$ | The transmission power density at user $i$ under Wi-Fi. |
| $Q_i^L$ | The transmission power density at user $i$ under LTE. |
| $\lambda_{ij}$ | The antenna gain between user $i$ and its service provider $j$ (either Wi-Fi or LTE) |
| $d_{ij}$ | The distance between user $i$ and its service provider $j$ (either Wi-Fi or LTE) |
| $\sigma$ | Path loss index. |

LTE and denote $P_i$ as user $i$'s ($i \in \mathcal{N}$) affordability, i.e., the maximum payment that user $i$ is willing to pay. When $P_i$ is 0, then user $i$ is not willing to pay and only wants to use free Wi-Fi service. Otherwise, user $i$ can get up to $P_i/p$ amount of data rate if she chooses LTE. Note that LTE provides guaranteed data rate while Wi-Fi only provides average rate (based on contention) which is likely to fluctuate over time. So, even for the same "rate", user experience under LTE and Wi-Fi will differ. To capture such difference in a user's experience, we introduce two satisfaction parameters for rates under LTE and Wi-Fi. We denote $S_W$ and $S_L$ as the satisfaction parameters per unit of data rate under LTE and Wi-Fi, respectively. Table 7.1 lists notation in this chapter.

Based on this setting, we are interesting in total users satisfaction under the following coexistence and spectrum-sharing strategies.

- (a) Wi-Fi only: Only Wi-Fi is deployed and the entire unlicensed spectrum is used by Wi-Fi. In this case, each user is served by one Wi-Fi AP.
- (b) Static partitioning of unlicensed spectrum between Wi-Fi and LTE: Both LTE and Wi-Fi are deployed in the area. The unlicensed band is partitioned into two fixed portions: one for Wi-Fi and the other for LTE. A user may be served by either a Wi-Fi AP or LTE BS. That is one of the coexistence strategies advocated by cellular carriers for sharing unlicensed spectrum with Wi-Fi.
- (c) Adaptive spectrum partitioning of unlicensed spectrum between Wi-Fi and LTE: Both LTE and Wi-Fi are deployed in the area. The unlicensed spectrum band is dynamically partitioned between Wi-Fi and LTE (no fixed allocation on unlicensed band) based on current user population and their affordabilities.

## 7.3 Scenario A: Wi-Fi Only

In this section, we consider the scenario where only Wi-Fi APs are deployed in the area and LTE is not deployed (i.e., Wi-Fi only). For this scenario, we develop the mathematical model and problem

formulation to maximize total user satisfaction. For any user, we assume she is under the coverage of at least one Wi-Fi AP. Due to overlapping of coverage areas, a user may also be in the service area of multiple APs. To model which AP is selected by a user, denote binary variable $x_{ij}$ as whether user $i \in \mathcal{N}$ selects Wi-Fi AP $j, j \in \mathcal{A}_i$, i.e.,

$$x_{ij} = \begin{cases} 1 & \text{If user } i \text{ selects Wi-Fi AP } j \text{ as her service provider;} \\ 0 & \text{otherwise.} \end{cases} \tag{7.3.1}$$

Since user $i$ can only select one and only one Wi-Fi AP, we have:

$$\sum_{j \in \mathcal{A}_i} x_{ij} = 1. \tag{7.3.2}$$

Since uplink and downlink traffic behavior is highly unpredictable, to simplify our study, we assume saturated traffic for each user. Also, since there does not exist a good throughput model that considers both uplink and downlink traffic for a user in Wi-Fi, we will only consider uplink traffic in this study and defer the more complex (unknown) joint uplink/downlink traffic model to future study. Such simplification allows us to employ the empirical throughout model in [7,53]. On the unlicensed bandwidth $B$, each user needs to contend with other users to access this bandwidth. Under saturated user traffic model, air time is shared equally among all users [7, 53]. Recall that $\mathcal{N}_i$ is the set of users that are within the CSMA contention range of user $i$. Then user $i$ needs to contend with all these users in $\mathcal{N}_i$ to access the same channel. The transmission opportunity for user $i$ is therefore $\frac{1}{|\mathcal{N}_i|+1}$, i.e., air time is shared equally among the $(|\mathcal{N}_i| + 1)$ users. Denote $r_{ij}^W$ as the achievable uplink throughput for user $i$ when it selects AP $j$. Then the achievable uplink throughput for user $i$ can be expressed as following:

$$r_{ij}^W = \frac{\alpha}{|\mathcal{N}_i| + 1} B \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma} \lambda_{ij}}{N_0}), \tag{7.3.3}$$

where $\alpha$ is the channel efficiency of air time [7, 53], $Q_i^W$ is user $i$' power spectral density under Wi-Fi, $d_{ij}$ is the distance between user $i$ and AP $j$, $\sigma$ is the pass loth index, $\lambda_{ij}$ is the antenna gain between user $i$ and AP $j$, and $N_0$ is the ambient Gaussian power spectral density.

Note the throughput in Eq. (7.3.3) is average (contention-based) throughput and the instantaneous rate will fluctuate over time. Recall $S_W$ is the satisfaction parameter per unit of data rate

under Wi-Fi. To capture a user's satisfaction, we define $f(i)$ as user $i$'s satisfaction function as follows:

$$f(i) = S_W \cdot \sum_{j \in \mathcal{A}_i} x_{ij} r_{ij}^W. \tag{7.3.4}$$

We are interesting in maximizing the total users satisfaction in the network. That is:

OPT-W

$$\begin{aligned} \max \quad & \textstyle\sum_{i \in \mathcal{N}} f(i) \\ \text{s.t.} \quad & \text{Satisfaction function: (7.3.4);} \\ & \text{AP selection constraints: (7.3.2);} \\ & \text{Throughput constraints: (7.3.3).} \end{aligned}$$

This problem is in the form of a mixed-integer linear program (MILP), which can be solved by commercial solver (CPLEX) efficiently.

## 7.4 Scenario B: Coexistence Through Static Spectrum Partitioning

### 7.4.1 Mathematical Modeling

In this deployment scenario, both Wi-Fi APs and LTE are deployed in the area (Fig. 7.1). Under static spectrum partitioning, Wi-Fi and LTE will coexist on the same unlicensed band $B$ and the total bandwidth $B$ is statically partitioned into $B_W$ and $B_L$ for Wi-Fi and LTE, respectively and remain fixed. To avoid interference between Wi-Fi and LTE, there is no overlap between $B_W$ and $B_L$.

**Service selection** A user may choose a Wi-Fi AP or LTE BS. The binary variable $x_{ij}$ (defined in (7.3.1)) can be used as an indicator of whether user $i$ selects AP $j$. Now denote $x_{iL}$ as a binary

variable indicating whether or nor user $i$ selects LTE BS as its service provider, i.e.,

$$x_{iL} = \begin{cases} 1 & \text{If user } i \text{ selects LTE BS as her service provider;} \\ 0 & \text{otherwise.} \end{cases}$$

Since a user can be served by either the LTE BS or one (and only one) Wi-Fi AP, we have:

$$x_{iL} + \sum_{j \in \mathcal{A}_i} x_{ij} = 1, \qquad (i \in \mathcal{N}). \tag{7.4.1}$$

**Bandwidth Allocation for LTE User** LTE BS typically has advanced channel management function and can slice its bandwidth $B_L$ into a set of different (and smaller) channels to serve its users. Denote $B_i^L$ as the bandwidth allocated to user $i$ by the LTE BS. To avoid potential interference among users in the LTE network, the channels assigned to different users should not overlap. That is:

$$\sum_{i \in \mathcal{N},\ x_{iL}=1} B_i^L \leq B_L.$$

which is equivalent to:

$$\sum_{i \in \mathcal{N}} x_{iL} B_i^L \leq B_L. \tag{7.4.2}$$

We define $B_{\min}^L$ as the minimum bandwidth that should be assigned to a user if it is served by LTE BS. If $x_{iL} = 1$, then $B_i^L \geq B_{\min}^L$; otherwise, $B_i^L = 0$. That is:

$$x_{iL} B_{\min}^L \leq B_i^L \leq x_{iL} B_L. \tag{7.4.3}$$

**Throughput Analysis** We now analyze a user's throughput. As for the Wi-Fi only network in Section 7.3, we only consider uplink traffic.

- **User $i$ served by Wi-Fi network.** For a user $i$ that is serviced by the Wi-Fi network, it contends the channel access with other Wi-Fi users in $\mathcal{N}_i$. Since the set $\mathcal{N}_i$ includes all users (using either Wi-Fi or LTE service) that are within the CSMA contention range of user $i$,

we need to identify only those users in $\mathcal{N}_i$ that are using Wi-Fi. Denote $M_i$ as the number of users in $\mathcal{N}_i$ that are served by Wi-Fi. Then user $i$ only contends with $M_i$ Wi-Fi users for channel $B_W$ that is allocated to Wi-Fi. $M_i$ can be modeled as following:

$$M_i = \sum_{k \in \mathcal{N}_i} \sum_{a \in \mathcal{A}_k} x_{ka}, \qquad (i \in \mathcal{N}). \tag{7.4.4}$$

If user $i$ selects Wi-Fi AP $j$, then based on our earlier discussion in Section 7.3, the achievable uplink throughput $r_{ij}^W$ is :

$$r_{ij}^W = \frac{\alpha}{M_i + 1} B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma} \lambda_{ij}}{N_0}). \tag{7.4.5}$$

- **User $i$ served by LTE network.** If user $i$ selects the LTE BS as its service provider, then LTE BS will assign a dedicated channel $B_i^L$ to it. Denote $r_i^L$ as the achievable uplink throughput for user $i$ under LTE. We have:

$$r_i^L = B_i^L \log_2(1 + \frac{Q_i^L d_{iL}^{-\sigma} \lambda_{iL}}{N_0}), \tag{7.4.6}$$

where $Q_i^L$ is user $i$' power spectral density under LTE, $d_{iL}$ is the distance between user $i$ and LTE BS, $\sigma$ is the pass loth index, $\lambda_{iL}$ is the antenna gain between user $i$ and LTE BS, and $N_0$ is the ambient Gaussian power spectral density.

**User Affordability Constraint** Recall that a user will need to pay for accessing LTE service. To characterize the financial means (affordability) of a user, we employ the following pricing model. Recall that we have defined $p$ as the price per unit of data rate imposed by LTE and $P_i$ as the upper limit that user $i$ is willing to pay. If a user chooses LTE, we have the following constraint:

$$p \cdot r_i^L \leq P_i. \tag{7.4.7}$$

### 7.4.2 Problem Formulation

Recall that the throughput in (7.4.6) for LTE is a guaranteed rate while the throughput in (7.4.5) is the average (contention-based) throughput. As a result, even for the same "rate", user $i$'s experience under LTE and Wi-Fi will differ. To capture such difference in user $i$'s satisfaction, we

introduce another satisfaction parameter for the user's rate under LTE. Denote $S_L$ as the satisfaction parameter per unit of data rate under LTE. Recall that $S_W$ is the satisfaction parameter per unit of data rate under Wi-Fi service. Therefore, for practical purpose, we should have $S_L \geq S_W$. Based on (7.3.4), we define $f(i)$ as the user $i$'s satisfaction function as follows:

$$f(i) = \begin{cases} S_W \cdot \sum_{j \in \mathcal{A}_i} x_{ij} r_{ij}^W & \text{If } \sum_{j \in \mathcal{A}_i} x_{ij} = 1; \\ S_L \cdot x_{iL} \cdot r_i^L & \text{if } x_{iL} = 1. \end{cases}$$

Since $x_{iL} + \sum_{j \in \mathcal{A}_i} x_{ij} = 1$, it is easy to show that the above definition of $f(i)$ is equivalent to:

$$f(i) = S_W \sum_{j \in \mathcal{A}_i} x_{ij} r_{ij}^W + S_L x_{iL} \cdot r_i^L, \qquad (i \in \mathcal{N}). \tag{7.4.8}$$

For the objective of maximizing total satisfaction among all users, we can formulate the problem as follows:

$$\begin{aligned} \max \quad & \textstyle\sum_{i \in \mathcal{N}} f(i) \\ \text{s.t.} \quad & \text{Satisfaction function: (7.4.8);} \\ & \text{Service selection constraints: (7.4.1);} \\ & \text{Bandwidth allocation constraints: (7.4.2), (7.4.3);} \\ & \text{Throughput constraints: (7.4.4), (7.4.5), (7.4.6);} \\ & \text{User affordability constraint: (7.4.7).} \end{aligned}$$

In this formulation, $x_{ij}, x_{iL}, M_i, B_i^L, r_{ij}^W$, and $r_i^L$ are optimization variables, and $B_W$, $B_L$, $B$, $B_{\min}^L$, $B_{\min}^W$, $p$, $P_i$, $S_W$, $S_L$, and $\alpha$ are constants. This optimization is in the form of a mixed-integer nonlinear program (MINLP). In the following, we show how to reformulate it into an MILP problem, which could be solved by a commercial software (such as CPLEX).

### 7.4.3 Reformulation

In above formulation, constraints (7.4.2), (7.4.5), and (7.4.8) are nonlinear. We will linearize them into a set of linear constraints.

In constraints (7.4.2) and (7.4.8), we have nonlinear terms $x_{iL}B_i^L$, $x_{ij}r_{ij}^W$, and $x_{iL}r_i^L$. We can use *Reformulation-Linearization technique (RLT)* [27, Chapter 6] to linearize such product of variables (monomials). Define $z_{iL} = x_{iL}B_i^L$, we have the following associate constraints:

$$x_{iL} \geq 0, \quad 1 - x_{iL} \geq 0.$$

$$B_i^L \geq 0, \quad B_L - B_i^L \geq 0.$$

We can cross-multiply the two constraints involving $x_{iL}$ with the two constraints involving $B_i^L$, and replacing the product term $(x_{iL}B_i^L)$ with $z_{iL}$. Then (7.4.2) can be replaced by the following linear constraints:

$$\sum_{i \in \mathcal{N}} z_{iL} \leq B_L, \tag{7.4.9}$$

$$z_{iL} \leq x_{iL}B, \tag{7.4.10}$$

$$z_{iL} \leq B_i^L, \tag{7.4.11}$$

$$z_{iL} \geq x_{iL}B + B_i^L - B_L, \tag{7.4.12}$$

where $i \in \mathcal{N}$.

Following the same token, define $\mu_{ij} = x_{ij}r_{ij}^W$ and $\theta_i = x_{iL}r_i^L$, we have the following associate constraints:

$$x_{ij} \geq 0, \;\; 1 - x_{ij} \geq 0, \;\; r_{ij}^W \geq 0, \;\; \alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma} \lambda_{ij}}{N_0}) - r_{ij}^W \geq 0.$$

$$x_{iL} \geq 0, \;\; 1 - x_{iL}, \;\; r_i^L \geq 0, \;\; B_L \log_2(1 + \frac{Q_i^L d_{iL}^{-\sigma} \lambda_{iL}}{N_0}) - r_i^L \geq 0.$$

We can cross-multiply the constraints involving $x_{ij}$ with the two constraints involving $r_{ij}^W$ and cross-multiply the constraints involving $x_{iL}$ with the two constraints involving $r_i^L$, and replacing the product terms $(x_{ij}r_{ij}^W)$ and $(x_{iL}r_i^L)$ with $\mu_{ij}$ and $\theta_i$. Then, (7.4.8) can be replaced by the following constraints:

$$f(i) = S_W \sum_{j \in \mathcal{A}_i} \mu_{ij} + S_L \theta_i, \tag{7.4.13}$$

$$\mu_{ij} \leq r_{ij}^W, \tag{7.4.14}$$

$$\mu_{ij} \leq x_{ij}\alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{ij}}{N_0}), \tag{7.4.15}$$

$$\mu_{ij} \geq r_{ij}^W + x_{ij}\alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{ij}}{N_0}) - \alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{ij}}{N_0}), \tag{7.4.16}$$

$$\theta_i \leq r_i^L, \tag{7.4.17}$$

$$\theta_i \leq x_{iL} B_L \log_2(1 + \frac{Q_i^L d_{iL}^{-\sigma}\lambda_{iL}}{N_0}), \tag{7.4.18}$$

$$\theta_i \geq r_i^L + x_{iL} B_L \log_2(1 + \frac{Q_i^L d_{iL}^{-\sigma}\lambda_{iL}}{N_0}) - B_L \log_2(1 + \frac{Q_i^L d_{iL}^{-\sigma}\lambda_{iL}}{N_0}). \tag{7.4.19}$$

where $i \in \mathcal{N}$.

Constraint (7.4.5) can be written in the following form:

$$M_i r_{ij}^W + r_i^W = \alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{ij}}{N_0})$$

Since $M_i r_{ij}^W = \sum_{k\in\mathcal{N}_i}\sum_{a\in\mathcal{A}_k} x_{ka} r_{ij}^W$, define $\lambda_{i,k,a,j} = x_{ka} r_{ij}^W$, we have the following associate constraints:

$$x_{ka} \geq 0,\ \ 1 - x_{ka} \geq 0,\ \ r_{ij}^W \geq 0,\ \ \alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{ij}}{N_0}) - r_{ij}^W \geq 0. \tag{7.4.20}$$

We can cross-multiply the constraints involving $x_{ka}$ with the two constraints involving $r_{ij}^W$, and replacing the product term $(x_{ka} r_{ij}^W)$ with $\lambda_{i,k,a,j}$. Then (7.4.5) can be replaced by the following linear constraints:

$$\sum_{k\in\mathcal{N}_i}\sum_{a\in\mathcal{A}_k} \lambda_{i,k,a,j} + r_i^W = \alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{iW}}{N_0}), \tag{7.4.21}$$

$$\lambda_{i,k,a,j} \leq r_{ij}^W, \tag{7.4.22}$$

$$\lambda_{i,k,a,j} \leq x_{ka}\alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{iW}}{N_0}), \tag{7.4.23}$$

$$\lambda_{i,k,a,j} \geq r_{ij}^W + x_{ka}\alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{ij}}{N_0}) - \alpha B_W \log_2(1 + \frac{Q_i^W d_{ij}^{-\sigma}\lambda_{ij}}{N_0}). \tag{7.4.24}$$

where $i \in \mathcal{N}, j \in \mathcal{A}, k \in \mathcal{N}_i$, and $a \in \mathcal{A}_k$.

Now, all nonlinear constraints in the original formulation are linear. We have the following new formulation:

OPT-S

max $\sum_{i\in\mathcal{N}} F(i)$

s.t. Satisfaction function: (7.4.13)–(7.4.19);

Service selection constraints: (7.4.1);

Bandwidth allocation constraints: (7.4.3), (7.4.9)–(7.4.12);

Throughput constraints: (7.4.4), (7.4.6), (7.4.21)–(7.4.24);

User affordability constraint: (7.4.7).

This formulation is in the form of mix-integer linear program (MILP), which can be solved by commercial software (CPLEX).

## 7.5 Scenario C: Coexistence Through Adaptive Spectrum Partitioning

Since the cloud server can perform centralized optimization, it is possible to share the unlicensed spectrum dynamically between Wi-Fi and LTE based on the users in the network. That is, $B_W$ and $B_L$ can be optimization variables rather than pre-assigned constants.

Since $B$ is partitioned into $B_W$ for Wi-Fi and $B_L$ for LTE, and there is no overlap between the two, we have:

$$B_W + B_L = B. \tag{7.5.1}$$

Here $B_W$ and $B_L$ are variables, and could be dynamically adjusted based on the current user population in the network.

Different from Eq. (7.4.2), there is no need to allocate extra bandwidth to LTE users beyond their requirement. So the constraint in Eq. (7.4.2) should be binding instead of an upper bound. We have:

$$\sum_{i\in\mathcal{N}} x_{iL} B_i^L = B_L. \tag{7.5.2}$$

Therefore, any bandwidth unused by LTE will be allocated to Wi-Fi users.

To ensure there is some minimum bandwidth for Wi-Fi users, denote $B_{\min}$ as the minimum bandwidth that is guaranteed for Wi-Fi. Then, we have

$$B_W \geq B_{\min}^W. \tag{7.5.3}$$

If a user is served by LTE, it has a minimum bandwidth for $B_i^L$, we have:

$$x_{iL} B_{\min}^L \leq B_i^L \leq x_{iL} B_L. \tag{7.5.4}$$

Then the objective of total users' satisfaction can be maximized with the following problem formulation:

OPT-D

max $\sum_{i \in \mathcal{N}} f(i)$

s.t. Satisfaction function: (7.4.8);

Service selection constraints: (7.4.1);

Spectrum partitioning constraint: (7.5.1);

Bandwidth allocation constraints: (7.5.2), (7.5.3), (7.5.4);

Throughput constraints: (7.4.4), (7.4.5), (7.4.6);

User affordability constraint: (7.4.7).

In this formulation, $x_{ij}, x_{iL}, M_i, B_W, B_L, B_i^L, r_{ij}^W$, and $r_i^L$ are optimization variables, and $\alpha$, $B$, $B_{\min}^L$, $B_{\min}^W$, $p$, $P_i$, $S_W$, and $S_L$ are constants. This optimization problem is in the form of a mixed-integer nonlinear program (MINLP). Again, we can use the similar approaches as in Section 7.4.3 to linearize the nonlinear constraints. Then, the reformulated problem becomes an MILP.

Table 7.2 lists the constants and optimization variables in the formulation of three deployment scenarios.

Table 7.2: The constants and optimization variables in the formulation of Wi-Fi only, static spectrum partition, and adaptive spectrum partition.

| | Wi-Fi only | Static Spectrum Partitioning | Adaptive Spectrum Partitioning |
|---|---|---|---|
| Constants | $\alpha, S_W, B, B^W_{\min}$ | $\alpha, B_W, B_L, B, B^L_{\min}, B^W_{\min}, p, P_i, S_W, S_L$ | $\alpha, B, B^L_{\min}, B^W_{\min}, p, P_i, S_W, S_L$ |
| Optimization Variables | $x_{ij}, r^W_{ij}$ | $x_{ij}, x_{iL}, M_i, B^L_i, r^W_{ij}, r^L_i$ | $x_{ij}, x_{iL}, M_i, B^L_i, B_W, B_L, r^W_{ij}, r^L_i$ |

## 7.6 Performance Evaluation

In this section, we perform extensive simulation studies to compare maximum users satisfaction objectives under the three spectrum usage strategies. Our findings are rather interesting. First, in terms of maximizing total user satisfaction function, we find that there does not appear to be any advantage of coexistence between Wi-Fi and LTE with static spectrum partitioning (when compared to Wi-Fi only scheme). This is interesting as it suggests that one might just deploy Wi-Fi without LTE in the unlicensed spectrum. This finding serves as a powerful counter argument to some telecom carriers' proposals to partition the unlicensed spectrum statically between Wi-Fi and LTE. Another finding shows that coexistence between Wi-Fi and LTE is only meaningful (or beneficial) if spectrum is partitioned in an adaptive manner. Otherwise, it may not be worth doing any coexistence at all.

### 7.6.1 Parameter Setting

We consider one LTE BS and multiple Wi-Fi APs that are randomly deployed in a circular area with radius 100. The LTE BS is at the center of the circle (see Figure 7.3). For generality, we normalize units for distance, bandwidth, power, data rate, and pricing with appropriate dimensions. We assume LTE BS and Wi-Fi APs' have coverage radii (transmission range) of 100 and 40, respectively. The CSMA contention (interference) range for Wi-Fi is 70. The total bandwidth that is available in the unlicensed spectrum is $B = 100$. The minimum bandwidth reserved for Wi-Fi network is $B_{\min} = 10$ (under coexistence with LTE). The transmission power spectrum density for

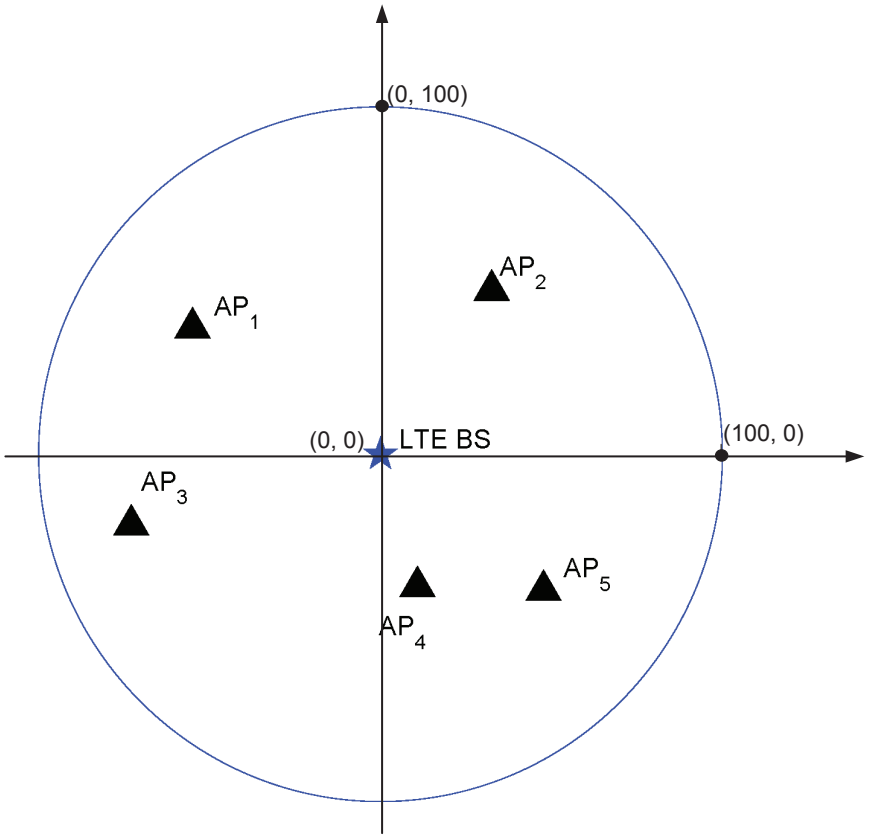


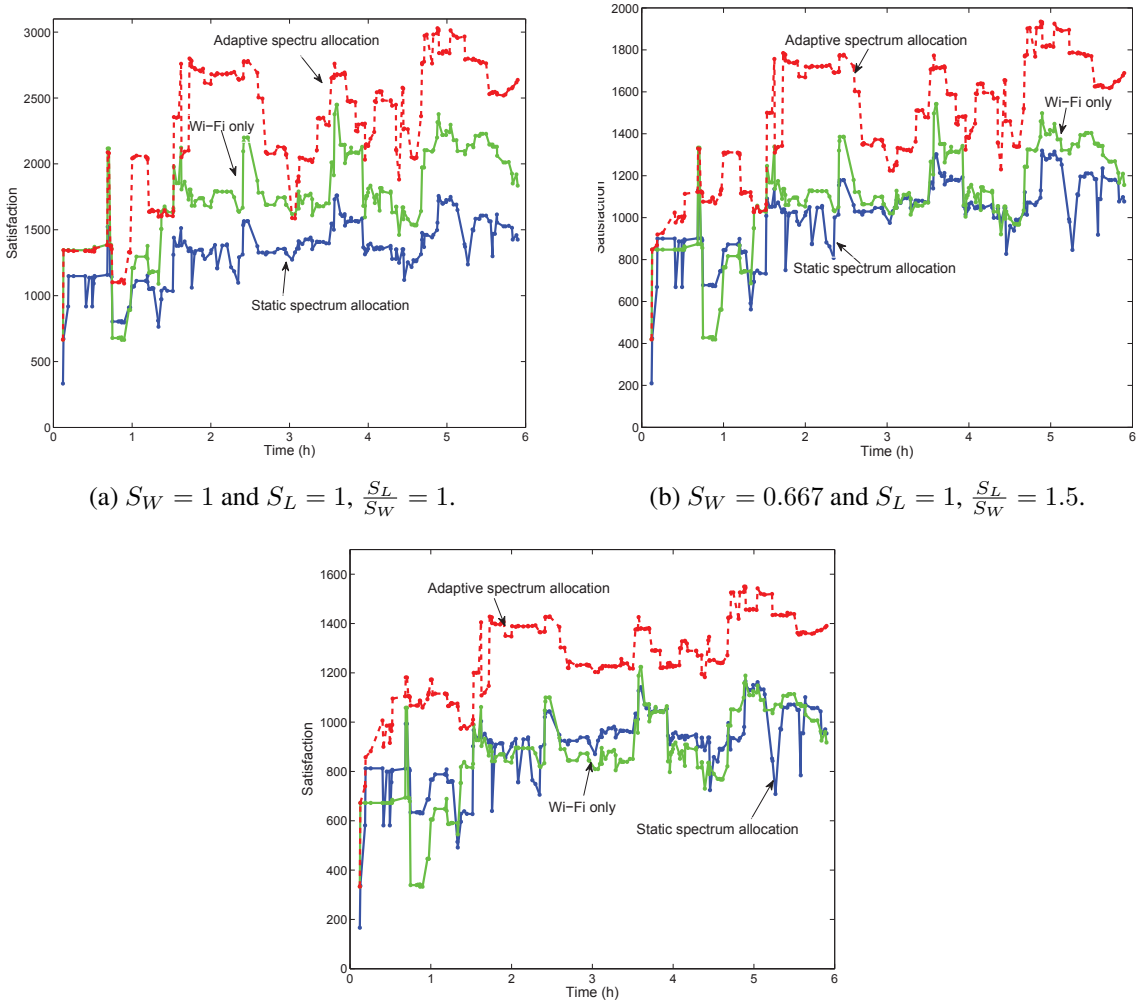

Figure 7.3: One LTE BS and multiple Wi-Fi APs that are randomly deployed in a circle with radius 100.

each user under Wi-Fi and LTE are 1.0 and 3.0, respectively. The ambient Gaussian power spectral density is $N_0 = 10^{-6}$. The path loss $\sigma$ is 3. The antenna gains are 1 between user and Wi-Fi AP and 2 between the user and LTE BS. We assume channel efficiency for Wi-Fi is $\alpha = 70\%$ [7]. Assume the price per unit of data rate charged by LTE is $p = 0.1$. For each user, her affordability is generated randomly. The user satisfaction coefficients for Wi-Fi and LTE will be specified in the respective performance studies.

### 7.6.2 Comparison Under Different Satisfaction Coefficients

We assume users' request arrivals follow a Poisson process with a rate of 20 per hour and the holding time for each user session is exponentially distributed with a mean of 1 hour. Upon arrival, the user's location may be anywhere (randomly distributed) inside the circular area. The simulation time is 6 hours. We perform simulation studies under various satisfaction parameters. We set the satisfaction parameter $S_L = 1$ and vary $S_W$ to 1, 0.67, and 0.5, respectively. That is, the ratios of satisfaction coefficients between LTE and Wi-Fi, $\frac{S_L}{S_W}$, are 1, 1.5, and 2, respectively. We compare the maximum user satisfaction objective values under Wi-Fi only (no LTE), coexistence between Wi-Fi and LTE with static spectrum partitioning, and coexistence between Wi-Fi and LTE with adaptive spectrum partitioning, respectively. Under static spectrum partitioning, we set $B_W = 50$ and $B_L = 50$.

Figs. 7.4(a), (b), and (c) show the maximum users satisfactions under different satisfaction parameters. We find that there is no advantage of coexistence between Wi-Fi and LTE with static spectrum partitioning over Wi-Fi only network. When $\frac{S_L}{S_W} = 1$ (Fig. 7.4(a)), the coexistence with static spectrum partitioning strategy performs even worse than Wi-Fi only. The reason is that when $\frac{S_L}{S_W} = 1$, for the same rate, there is no difference in terms of user satisfaction between Wi-Fi and LTE. On the other hand, static spectrum partitioning sets a hard partition between Wi-Fi and LTE. When bandwidth $B_L$ is not fully used, the remaining bandwidth still cantnot be used by Wi-Fi. Likewise when there is a need of more bandwidth for LTE users, Wi-Fi cannot release any bandwidth. When $\frac{S_L}{S_W} = 1.5$ and 2 (Figs. 7.4(b) and (c)), the satisfaction parameters favor

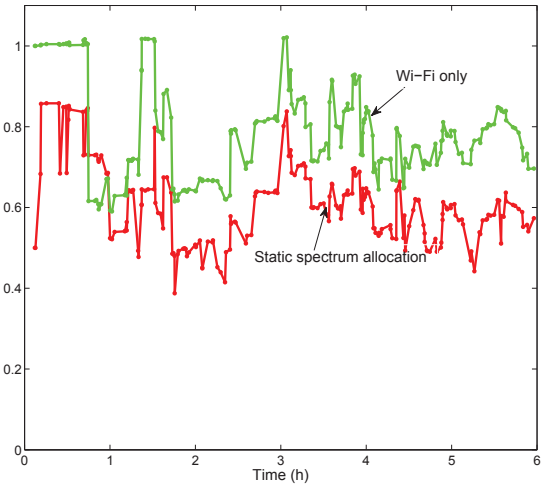



(a) $S_W = 1$ and $S_L = 1$, $\frac{S_L}{S_W} = 1$.

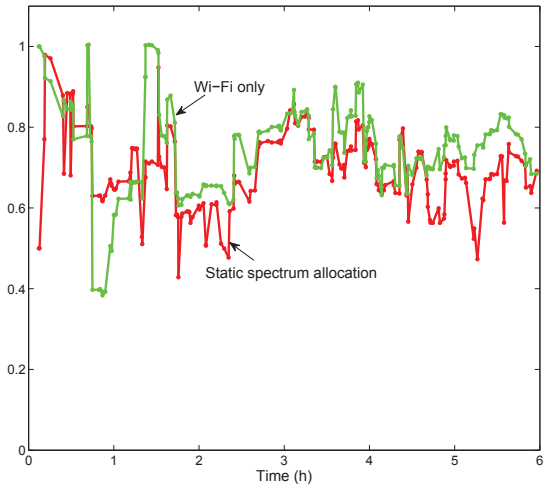
(b) $S_W = 0.667$ and $S_L = 1$, $\frac{S_L}{S_W} = 1.5$.

(c) $S_W = 0.5$, $S_L = 1$, $\frac{S_L}{S_W} = 2$.

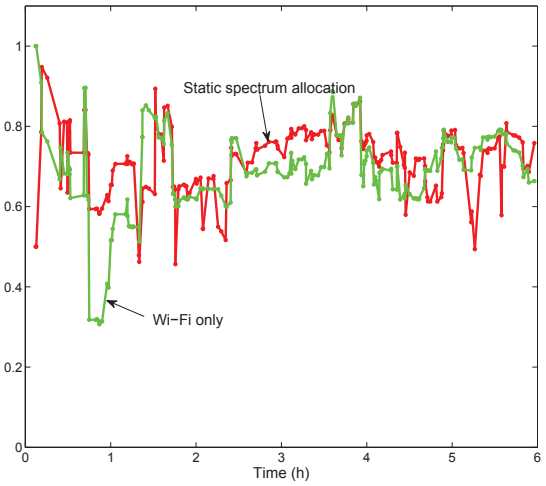
Figure 7.4: Maximum users satisfaction under Wi-Fi only, static spectrum partitioning, and adaptive spectrum partitioning with different satisfaction coefficients.




(a) $S_W = 1$ and $S_L = 1$, $\frac{S_L}{S_W} = 1$.




(b) $S_W = 0.667$ and $S_L = 1$, $\frac{S_L}{S_W} = 1.5$.




(c) $S_W = 0.5$, $S_L = 1$, $\frac{S_L}{S_W} = 2$.

Figure 7.5: Normalized users satisfaction of Wi-Fi only and static spectrum partitioning with respect to adaptive spectrum partitioning.

LTE network. But such favor still cannot overcome the adverse effective due to hard spectrum partitioning. In order words, the hard (static) partitioning between Wi-Fi and LTE has a much more significant impact than satisfaction parameter setting. Consequently, coexistence with static partitioning is not desirable for the goal of maximizing total users satisfaction.

On the other hand, we can see that the adaptive spectrum partitioning strategy always achieves the highest users satisfaction. To see the difference more clearly, in Figs. 7.5 (a), (b), and (c), we plot normalized users satisfaction for Wi-Fi only and static partitioning with respective to that for adaptive spectrum partitioning. In all cases, the ratio is less than 1, indicating adaptive spectrum partitioning has a dominant advantage over the other two.

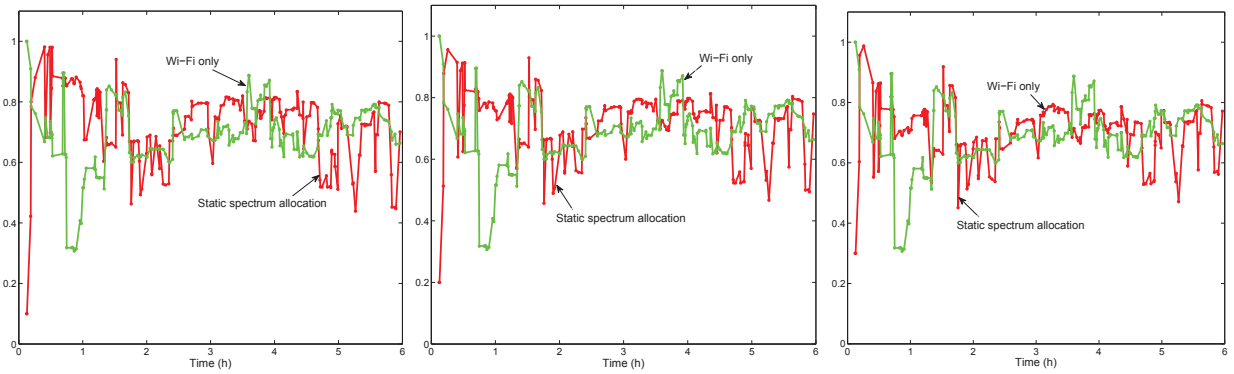
### 7.6.3 Different Bandwidth Allocation in Static Partitioning Scheme

In this study, we want to understand the impact of different bandwidth partitioning for $B_W$ and $B_L$ (under static spectrum partitioning) on maximum users satisfaction. We change $B_W$ from 10 to 90 (and correspondingly $B_L$ from 90 to 10). We set $S_L = 1$ and $\frac{S_L}{S_W} = 2$, which favors LTE. Figure 7.6 (a) to (i) show the normalized users satisfaction for Wi-Fi only and static spectrum partitioning with respective to those for adaptive spectrum partitioning. From these figures, we can see there is no clear benefits for coexistence between Wi-Fi and LTE with static spectrum partitioning over Wi-Fi only even when the user satisfaction parameters favor LTE. This further indicates that the adverse effect from static spectrum partitioning is very significant. On the other hand, coexistence under adaptive spectrum partitioning has a dominant advantage over the other two.

### 7.6.4 Varying Traffic Load

In this section, we compare maximum users satisfaction for the three strategies by varying traffic load. We set $S_W = 0.5$ and $S_L = 1$ (i.e., $\frac{S_L}{S_W} = 2$), which favors LTE for the same rate. Under static spectrum partitioning, we set $B_W = 50$ and $B_L = 50$.

Figures 7.7(a), (b), and (c) show the normalized users satisfaction for Wi-Fi only and static



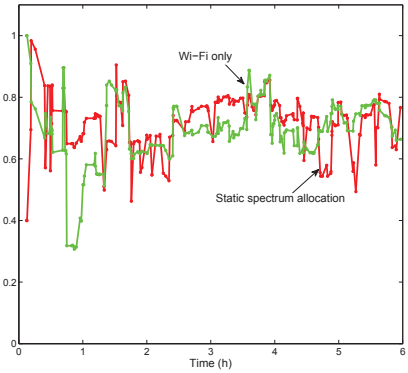

(a) $B_W = 10$ and $B_L = 90$.



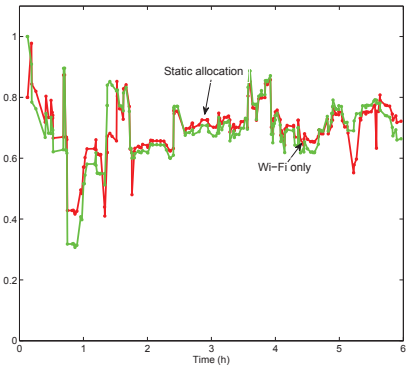

(b) $B_W = 20$ and $B_L = 80$.



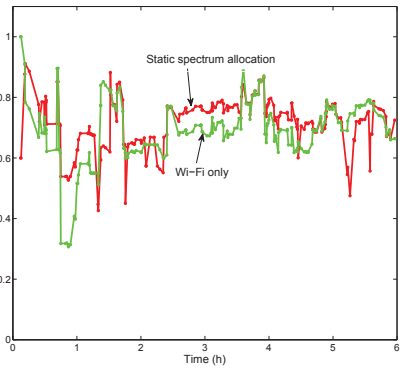

(c) $B_W = 30$ and $B_L = 70$.



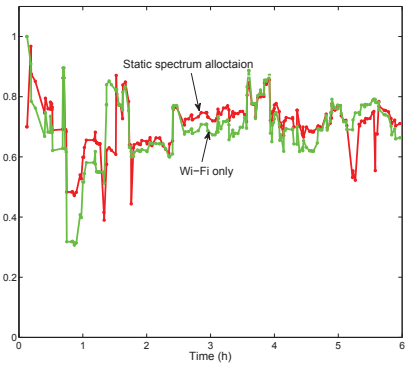

(d) $B_W = 40$ and $B_L = 60$.



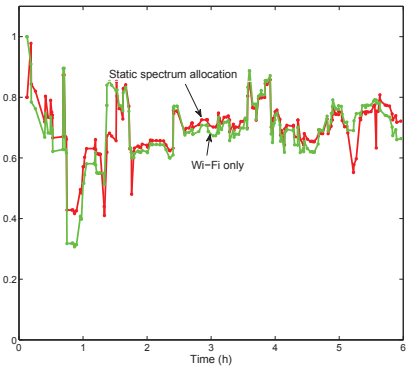

(e) $B_W = 50$ and $B_L = 50$.



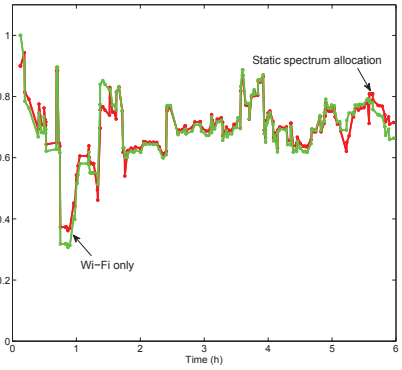

(f) $B_W = 60$ and $B_L = 40$.




(g) $B_W = 70$ and $B_L = 30$.

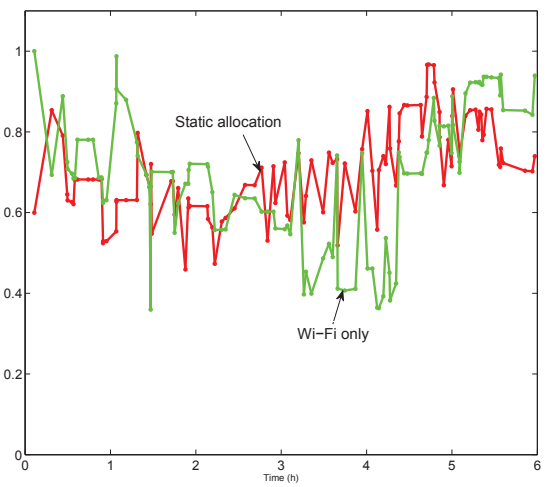



(h) $B_W = 80$ and $B_L = 20$.

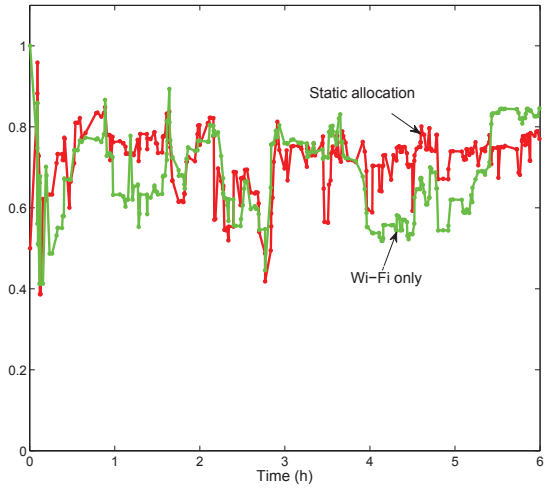



(i) $B_W = 90$ and $B_L = 10$.

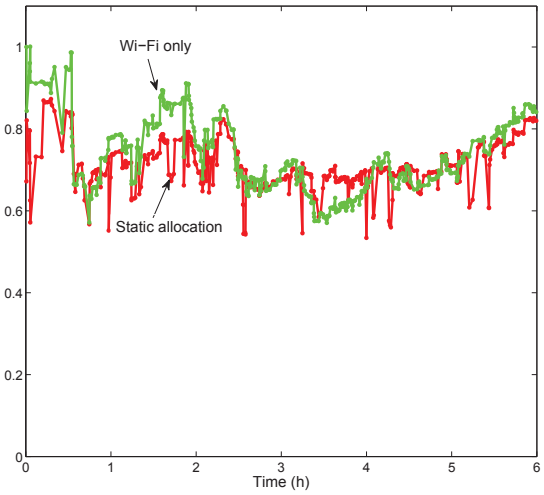
Figure 7.6: Normalized users satisfaction of Wi-Fi only and static spectrum partitioning under different bandwidth allocation with respect to those for adaptive spectrum partitioning.




(a) Users arrival rate is 10 per hour.




(b) Users arrival rate is 30 per hour.




(c) Users arrival rate is 50 per hour.

Figure 7.7: Normalized users satisfaction of Wi-Fi only and static spectrum partitioning with respect to those of adaptive spectrum partitioning when the user arrival rates are 10, 30, and 50 per hour.

spectrum partitioning with respect to those of adaptive spectrum partitioning when the user arrival rates are 10, 30, and 50 per hour. From these figures, we can see there is no clear benefits for coexistence between Wi-Fi and LTE with static spectrum partitioning over Wi-Fi even when user satisfaction parameters favor LTE and coexistence under adaptive spectrum partitioning has a dominant advantage over the other two.

## 7.7 Semi-Adaptive Algorithm for Practical Implementing

### 7.7.1 Motivation

Based on our findings in Section 7.6, we conclude that adaptive spectrum partitioning is the only viable approach for coexistence between Wi-Fi and LTE from user satisfaction perspective. But the adaptive spectrum partitioning scheme in Section 7.5 is based on global optimization across all users, meaning that $x_{iL}$, $x_{ij}$, $M_i$, $B_W$, $B_L$, $B_i^L$, $r_{ij}^W$, and $r_i^L$ are all optimization variables. This approach cannot be implemented in practice. This is because each time when there is a new request arrival (or a departure of an existing user), the centralized optimization will be executed and yield a new solution for all users. As a result, an existing user may need to change her current service provider (e.g., from Wi-Fi to LTE or vice versa, or switch to a different Wi-Fi AP). Such frequent change of service provider is quite disruptive at the application layer and should be avoided. What we need is a semi-adaptive algorithm that does not affect the service providers for existing users. In this section, we will design such a semi-adaptive algorithm, in which service providers for existing users will not change but only bandwidth partitioning and allocation may change.

### 7.7.2 Algorithm Design

**Roadmap** The design goal of our proposed algorithm is to optimally handle a new user request or departure of an existing user with minimum impact on existing users. Specially, under either event (arrival or departure), the service provider for any of the existing users should not be affected.

What can be changed for the existing users are the allocated bandwidth, i.e., $B_W$ for Wi-Fi users and $B_i^L$ for LTE users which can be adjusted rather easily based on today's programmable radio technologies.

When a new user arrives, it will send its request to the cloud server (via its neighboring Wi-Fi AP). Upon receiving this request, the cloud server will formulate a new users satisfaction problem by considering the service provider for existing users being fixed (pre-assigned) and only service provider for the new user and bandwidth allocation for all users being variables. After finding a new optimal solution, the cloud server sends bandwidth allocation to all users (via Wi-Fi APs and LTE BS) and service selection to the new user. Upon an existing user terminates, the user will send a termination message to the cloud server. Upon receiving this message, the cloud server will re-optimize bandwidth allocation for all users in both Wi-Fi and LTE.

Since the cloud server performs all computation for resource allocation, a set of information must be maintained at the cloud server. Specially, the following information should be maintained:

- Service Selection: The cloud server should maintain the service provider selection for each user, i.e, $x_{ij}$ and $x_{iL}$.
- Bandwidth Partitioning: The cloud server should maintain the bandwidth partition for the Wi-Fi network (i.e., $B_W$) and LTE network (i.e., $B_L$).
- Bandwidth Allocation: The cloud server should maintain bandwidth allocation for each user under LTE ($B_i^L$).

**Algorithm Details** Now, we present the details of our semi-adaptive algorithm when a user initiates and terminates its service.

- **Initiation of A New User.** When a new user initiates a request to access the network, it will send a control message to its neighboring Wi-Fi AP. The request message includes the users' affordability. The Wi-Fi AP sends the request message to the cloud server. Upon

receiving the request message, the cloud server solves the following optimization problem (OPT-Arrival), where $k$ denotes the new user.

OPT-Arrival

$$\begin{aligned} \max \quad & \textstyle\sum_{i\in\mathcal{N}\cup\{k\}} f(i) \\ \text{s.t.} \quad & \text{Satisfaction function (7.4.8) with } x_{ij}(i \in \mathcal{N}) \text{ being constants} \\ & \text{and } x_{kj} \text{ as variable;} \\ & \text{Service selection constraint only for new user } k\text{: } x_{kL} + \textstyle\sum_{j\in\mathcal{A}_k} x_{kj} = 1; \\ & \text{Spectrum partitioning constraint: (7.5.1);} \\ & \text{Bandwidth allocation constraints: (7.5.2), (7.5.3), (7.5.4);} \\ & \text{Throughput constraints: (7.4.4), (7.4.5), (7.4.6);} \\ & \text{User affordability constraint: (7.4.7).} \end{aligned}$$

In this formulation, $x_{kL}$, $x_{kj}$, $B_i^L$, $B_W$, $B_L$, $M_i$, $r_{ij}^W$, and $r_k^L$ are variables. $\mathcal{N}$ denotes the set of existing users in the network. $x_{ij}$ and $x_{iL}$ for existing users $i \in \mathcal{N}$ are constants. This optimization problem is in the form of a mixed-integer nonlinear program (MINLP). We can use the same RLT technique as in Section 7.4.3 to reformulate all nonlinear constraints into linear constraints and obtain an MILP, which can be solved by a commercial solver (CPLEX).

After finding a new solution, the cloud server stores the service selection variable $x_{kL}$ and update spectrum partitioning variables $B_W, B_L$, and bandwidth allocation variable $B_i^L$. Then it sends updates to all user via their Wi-Fi or LTE service providers. Based on new spectrum partitioning and bandwidth allocation information, each user's radio adjusts its operating bandwidth. The service providers for existing users are intact.

- **Termination of An Existing User.** When an existing user terminates its session, the user sends a termination message to the cloud server through its service provider. Upon receiving this termination message at the cloud server, it will remove user $k$ from $\mathcal{N}$, i.e., $\mathcal{N} = \mathcal{N}\backslash\{k\}$. Then it will formulate a user satisfaction problem to re-optimize spectrum partition and

the bandwidth allocation among the remaining users by solving the following optimization problem:

OPT-Departure

$$\begin{aligned} \max \quad & \textstyle\sum_{i\in\mathcal{N}} f(i) \\ \text{s.t.} \quad & \text{Satisfaction function: (7.4.8);} \\ & \text{Spectrum partitioning constraint: (7.5.1);} \\ & \text{Bandwidth allocation constraints: (7.5.2), (7.5.3), (7.5.4);} \\ & \text{Throughput constraints: (7.4.4), (7.4.5), (7.4.6);} \\ & \text{User affordability constraint: (7.4.7).} \end{aligned}$$

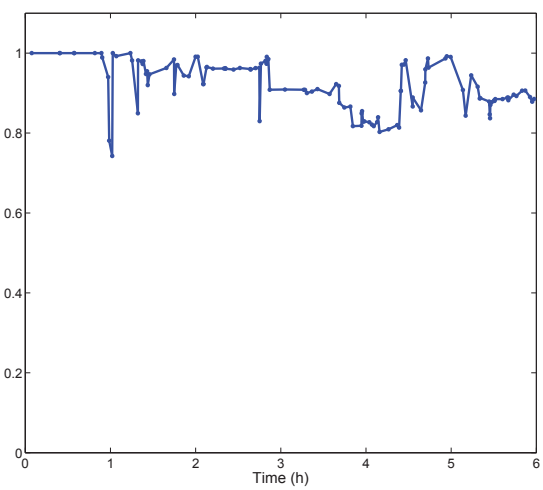
In this formulation, $B_i^L$, $B_W$, $B_L$, $r_{ij}^W$, and $r_i^L$ are variables, while $x_{ij}, x_{iL}$ and $M_i$ are constants. This problem is an MILP, which could be solved by CPLEX at cloud server.

After solving the optimization problem for spectrum partitioning and bandwidth allocation, the cloud server will send this update back to the users who will then adjust the bandwidths of their radios.

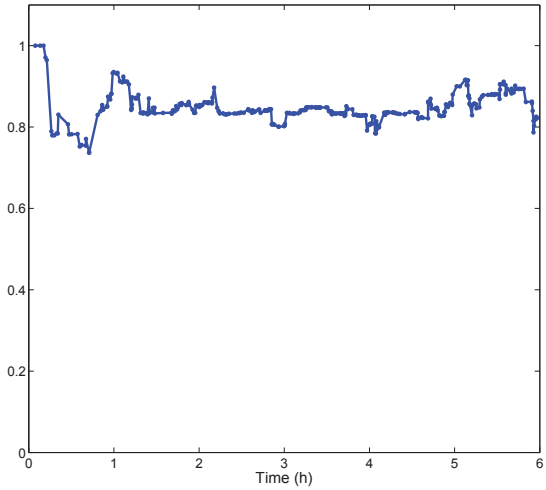
### 7.7.3 Performance Evaluation

Now we evaluate the performance of our proposed semi-adaptive algorithm. We use the same setting as in Section 7.6.1. We set the satisfaction coefficients to $S_W = 0.5$ and $S_L = 1$. We compare the objective values (maximum users satisfaction) from our proposed semi-adaptive algorithm to fully adaptive spectrum partitioning.

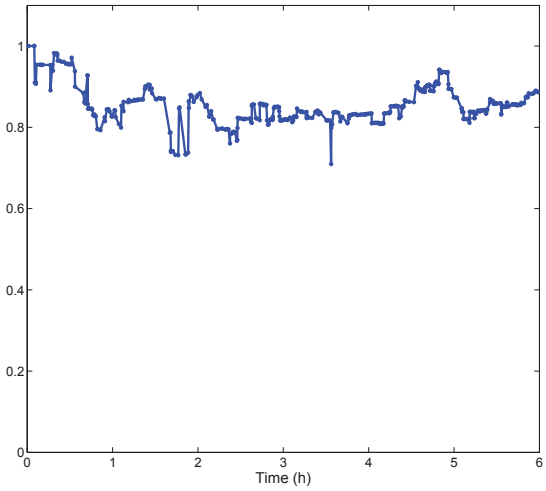
Figure 7.8(a), (b) and (c) show the normalized objective values from the semi-adaptive algorithm to the fully adaptive spectrum partitioning when the users arrival rates are 10, 30, and 50 per hour. In Figure 7.8(a), there are a total of 122 events during this simulation, among which there are 50 events with ratio over 95%, 75 events with ratio over 90%, 101 events with ratio over 85%, and 120 events with ratio over 80%. Figure 7.9(a) presents the CDF of the ratio. The average ratio between the two is 91.86%.




(a) Users arrival rate is 10 per hour.

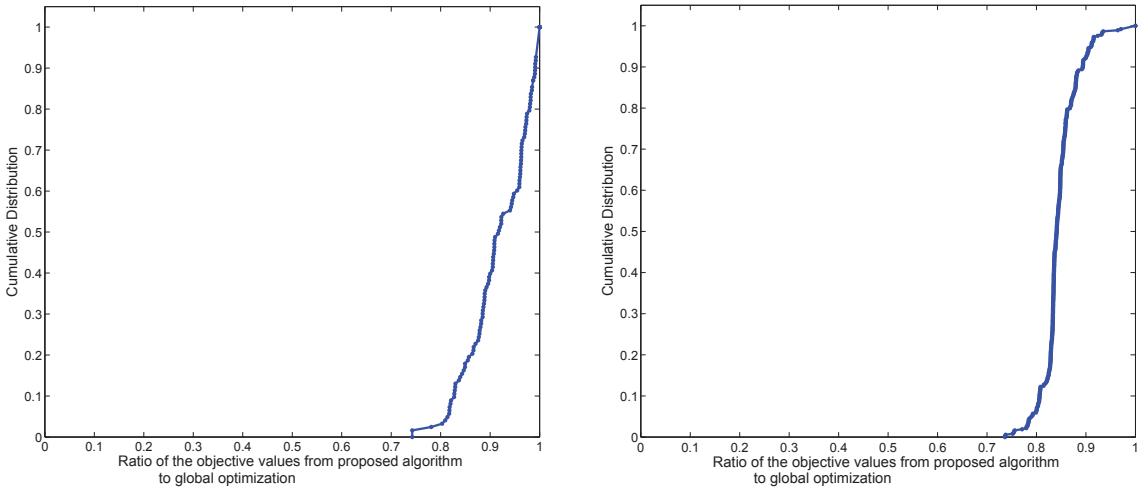



(b) Users arrival rate is 30 per hour.



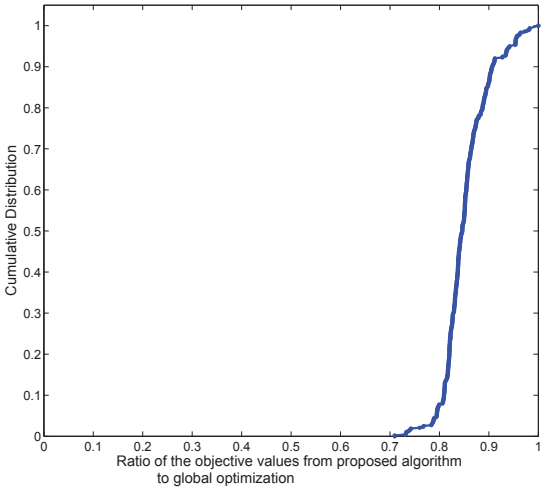

(c) Users arrival rate is 50 per hour.

Figure 7.8: Normalized objective value for the proposed semi-adaptive algorithm to fully adaptive spectrum partitioning with different user arrival rates.




(a) Users arrival rate is 10 per hour.

(b) Users arrival rate is 30 per hour.




(c) Users arrival rate is 50 per hour.

Figure 7.9: The CDFs of normalized objective values for the proposed semi-adaptive algorithm to fully adaptive spectrum partitioning under different user arrival rates.

In Figure 7.8(b), there are a total of 369 events, among which there are 30 events with ratio over 90%, 125 events with ratio over 85%, and 346 events with ratio over 80%. The CDF of the ratio is shown in Figure 7.9(b). The average ratio between the two is 84.6%.

In Figure 7.8 (c), there are a total of 482 events, among which there are 90 events with ratio over 90%, 197 events with ratio over 85%, and 385 events with ratio over 80%. The CDF of the ratio is shown in Figure 7.9(c). The average ratio between the two is 83.34%.

From the results in Figures 7.8 and 7.9, we can conclude that our proposed semi-adaptive algorithm is highly competitive when compared to fully adaptive spectrum partitioning.

Following the same validation methodology, we also run results with different network settings (i.e., network topology and satisfaction parameters). The results are consistent and show that our proposed algorithm is competitive.

## 7.8 Related Work

A number of approaches have been proposed to allow coexistence between LTE and Wi-Fi in the unlicensed bands. These approaches achieve coexistence between the two either in frequency domain or time domain.

In the frequency domain, coexistence between LTE-U and Wi-Fi can be achieved by having the two operate on separate, non-overlapping channels in the unlicensed band [54, 55]. This is called dynamic channel selection (DCS) in LTE-U. Under this approach, each channel consists of a 20 MHz band and Wi-Fi will use one of these bands that is not used by LTE-U. Given that there is no interference between Wi-Fi and LTE users after channel assignment, LTE users do not need to employ listen-before-talk (LBT). The biggest problem with this approach is that it follows the same traditional static spectrum partitioning on the unlicensed band. As a result, this approach will inherit all of the inefficiencies associated with traditional static spectrum partitioning, as we have demonstrated in this chapter.

In the time domain, when both Wi-Fi and LTE are using the same spectrum, one approach is to incorporate some form of LBT in LTE to make it compatible with Wi-Fi [36, 51, 64, 87]. This is known as carrier sensing adaptive transmission (CSAT) in LTE-U [47]. There are two issues with this approach. First, due to LBT, CSAT compromises the rate guarantee that users have been accustomed to under current LTE service. As a result, it is hard to justify why a user would choose LTE-U instead of using Wi-Fi directly, especially when Wi-Fi is increasingly being offered for free and a smartphone can easily switch to Wi-Fi. Second, CSAT may not be fair to Wi-Fi users, since the transmission period and resource allocation are solely controlled by LTE-U. Since CSAT may favor LTE-U over Wi-Fi, people in industry are skeptical about fairness for coexistence between the two technologies. Another approach is to mute or limit the transmission of LTE users such that LTE users access the channel in a fractional portion of air time. This is accomplished by the so-called Almost-Blank Subframes [1, 2, 24, 44, 88] or time partition for Wi-Fi and LTE [9, 11, 12, 38, 53]. The biggest problem with this approach is that it requires Wi-Fi to synchronize with LTE in order to access air time, which would involve a major change to the Wi-Fi protocol.

In addition to frequency and time domain coexistence, some approaches have employed physical layer techniques to achieve Wi-Fi/LTE coexistence (e.g., power control [10] and MIMO [80]).

## 7.9 Chapter Summary

This chapter studied different Wi-Fi and LTE deployment and coexistence scenarios from users satisfaction perspective. We investigate three scenarios, namely Wi-Fi only, static spectrum partitioning, and adaptive spectrum partitioning. We develop mathematical models and studied the problems of how to maximize total user satisfaction among all users under the three strategies. We find that in terms of maximizing total user satisfaction function, there does not appear to be any advantage with coexistence between Wi-Fi and LTE when the unlicensed spectrum is partitioned statically between Wi-Fi and LTE. This is interesting as it suggests that one might just deploy Wi-Fi without LTE in the unlicensed spectrum. This finding serves as a powerful counter argument to

some telecom carriers' proposal to statically partition unlicensed band between Wi-Fi and LTE. On the other hand, we find that there is significant advantage in deploying adaptive spectrum sharing (between Wi-Fi and LTE). This finding shows that a centralized coordinator is needed to dynamically partition bandwidth between Wi-Fi and LTE. Due to some practical issues in implementing fully adaptive spectrum sharing in practice, we proposed a semi-adaptive algorithm for practical implementation. Our performance evaluation show that the proposed semi-adaptive algorithm is highly competitive. The findings in the chapter shed new light on the current debate on coexistence between Wi-Fi and LTE and pointed out a new direction for future research in this area.

# Chapter 8

# Dissertation Summary and Future Work

## 8.1 Dissertation Summary

In this dissertation, we investigated novel spectrum sharing policies and coexistence mechanisms to enhance radio spectrum utilization. The work in this dissertation consists of three parts. In the first part, we studied the transparent coexistence paradigm to achieve simultaneously transmission between multi-hop primary and secondary networks in time, space, and frequency domains. This paradigm can remove the limitation of the interweave paradigm, where a secondary user can only exploit spectrum holes in time, space, and frequency domains. In Chapter 2, we presented new technical challenges to achieve the transparent coexistence paradigm in a multi-hop network. Through a rigorous modeling, problem formulation, solution development, and simulation results, we showed that transparent coexistence paradigm offers significant improvement in terms of spectrum access and throughput performance as compared to the interference avoidance paradigm. In Chapter 3, we designed a distributed iterative algorithm to achieve the transparent coexistence for multi-hop primary and secondary networks. We allowed each node only to maintain two local sets to keep track of its IC responsibilities. We showed how to establish, maintain, and update these two local sets for each node to ensure that IC is one efficiently and in a feasible manner. Although no explicit node ordering is maintained at each node, we showed that the use of two local sets can be

mapped to an explicit global node ordering for IC among all nodes in the network. This guaranteed that there exists a set of feasible precoding/decoding vectors at each node so that all data can be transported free of interference. In Chapter 4, we studied an online distributed algorithm to handle dynamic session arrival and departure in the transparent coexistence paradigm. We showed that our algorithm can ensure the transparent coexistence is achieved at all time under traffic dynamics (i.e., inter-network and intra-network IC are always feasible at the PHY layer at all time under traffic dynamics).

In the second part, we proposed a policy-based network cooperation paradigm. Under this paradigm, we considered the UPS policy as an example to show the advantages of network cooperations. In Chapter 5, we studied a problem with the goal of supporting the rate requirement of the primary traffic while maximizing the throughput of the secondary sessions. In Chapter 6, we offer an in-depth study of the UPS paradigm in term of the maximum achievable throughput for both primary and secondary users.

In the third part of this dissertation, we studied the coexistence of Wi-Fi and LTE on the unlicensed bands from users satisfaction perspective. We investigated three deployment and spectrum sharing strategies for Wi-Fi and LTE, namely Wi-Fi only, static spectrum partitioning, and adaptive spectrum partitioning. We developed mathematical models and studied the problems of how to maximize total user satisfaction among all users under the three strategies. We found that in terms of maximizing total user satisfaction function, there does not appear to be any advantage with coexistence between Wi-Fi and LTE when the unlicensed spectrum is partitioned statically between Wi-Fi and LTE. This is interesting as it suggests that one might just deploy Wi-Fi without LTE in the unlicensed spectrum, when the objective is to maximize total user satisfaction. This finding serves as a powerful counter argument to some of the current telecom carriers' proposal to statically partition unlicensed band between Wi-Fi and LTE. On the other hand, we find that there is significant advantage in deploying adaptive spectrum sharing (between Wi-Fi and LTE). This finding shows that a centralized coordinator is needed to dynamically partition bandwidth between Wi-Fi and LTE. Due to some technical issues in implementing adaptive spectrum sharing in practice, we proposed a semi-adaptive algorithm to implement adaptive spectrum partitioning.

## 8.2 Future Work

There is a wealth of opportunities for future research on enhancing utilization of radio spectrum. The following is a list of problems as a result of this dissertation. The list is not meant to be exhaustive, but only serves as an illustration of some possibilities.

- **Transparent coexistence paradigm.** Although we have shown the potential of transparent coexistence in terms of throughput improvement for the secondary networks, much work remains to be done to transition this idea into reality. We briefly discuss some of the practical issues that must be addressed in future work to achieve transparent coexistence in the real world. This discussion is not meant to be exhaustive, as the transparent coexistence is a novel concept and its path to adaptation is bound to encounter many challenges, both known and unknown. The first issue is that the secondary nodes need to have accurate knowledge of the primary nodes' transmission behavior (information regarding transmitter, receiver, time slot, and channel). This issue is easier to address in a single-hop environment (cellular, TV tower, WiFi) but is a major challenge in a multi-hop ad hoc network environment. Second, we assume the schemes in Section 2.2.1 to obtain CSI would work perfectly and channel reciprocity strictly holds. But in reality, the communication channel not only consists of the physical channel, but also the antennas, RF mixers, filters, A/D converters, etc., which are not necessarily identical on all the nodes. Therefore, complex calibration among the nodes is needed to achieve channel reciprocity. Such calibration is no simple task for a pair of transmitter and receiver and is even more complicated among a network of nodes. Third, zero-forcing based IC may not be perfect even if we have perfect CSI. A consequence of non-perfect IC is interference leakage, which is undesirable for both primary and secondary receivers. How to mitigate such interference leakage to a minimal acceptable level should be a key consideration when deploying transparent coexistence for real applications. Clearly, there is a large landscape for further research on these important practical operation issues.

- **Policy-based network cooperation paradigm.** In this report, we only exploited the UPS policy as an example to demonstrate the benefits of the policy-based network cooperation. In

our future work, we will explore other policies under the policy-based network cooperation paradigm. Under a given policy, data forwarding behavior may also be affected by user requirements and performance objectives. Such user requirements and performance objectives under a particular policy are many, and each scenario would result in different data forwarding for both the primary and the secondary sessions. Clearly, there is a large landscape for further research under this new paradigm. We hope our vision and results in this chapter will open the door for further research in this area.

- **Harmonizing competing Wi-Fi and cellular service providers.** As shown in Chapter 7, a harmonious coexistence between Wi-Fi and cellular can be achieved smoothly if both Wi-Fi and cellular service providers belong to the same telecom service provider. Since the objective there is to maximize total user satisfaction, the telecom service provider can rely on its central server to arbitrate spectrum allocation between Wi-Fi APs and its cellular base stations. But when Wi-Fi and cellular belong to different (competing) service providers, a harmonious coexistence between the two becomes more challenging. In this case, both Wi-Fi and cellular service provider that meets their financial means and service needs. An independent (third party) SAS would be necessary to serve as a spectrum broker on the unlicensed band between the two service providers. In this scenario, a mechanism is needed to ensure that neither Wi-Fi nor cellular will starve the other during the dynamic sharing of the unlicensed spectrum. To achieve this goal, the SAS will need more information from Wi-Fi and cellular providers beyond merely a request for spectrum. Specially, the SAS will need to know the total number of users requesting services and the breakdown of users serviced by Wi-Fi and cellular. Here, some meaningful objectives should be employed at SAS to optimally allocate spectrum between the two services. The definition of this objective function is important as its optimal solution will directly affect the spectrum allocation outcome. Wi-Fi and cellular service providers may also have their own objective functions. Therefore, we may need to study a multi-objective optimization involving objective functions for SAS, Wi-Fi, and cellular service providers.

# Bibliography

[1] F.M. Abinader, E.P.L. Almeida, F.S. Chaves, A.M. Cavalcante, R.D. Vieira, E.C.D. Paiva, A.M. Sobrinho, S. Choudhury, E. Tuomaala, K. Doppler, and V.A. Sousa, "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Communications Magazine,* vol. 55, no. 11, pp. 54–61, Nov. 2014.

[2] E. Almeida, A.M. Cavalcante, R.C.D. Paiva, F.S. Chaves, F.M. Abinader Jr., and R.D. Vieira, "Enabling LTE/WiFi coexistence by LTE blank subframe allocation," in *Proc. IEEE ICC,* pp. 5083–5088, Budapest, Hungary, June 9–13, 2013.

[3] I.F. Akyildiz, W.-Y. Lee, M.C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Elsevier Computer Networks,* vol. 50, no. 13, pp. 2127–2159, September 2006.

[4] I.F. Akyildiz, W.-Y. Lee, and K.R. Chowdhury, "CRAHNs: Cognitive radio ad hoc networks," *Elsevier Ad Hoc Networks,* vol. 7, no. 5, pp. 810–836, July 2009.

[5] O. Bakr, M. Johnson, R. Mudumbai, and K. Ramchandran, "Multi-antenna interference cancellation techniques for cognitive radio applications," in *Proc. IEEE WCNC,* Budapest, Hungary, 6-pages, April 2009.

[6] M.S. Bazaraa, J.J. Jarvis, and H.D. Sherali, *Linear Programming and Network Flows,* Fourth Edition, John Wiley & Sons, 2010.

[7] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications,* vol. 18, no. 3, pp. 535–547, March 2000.

[8] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H.V. Poor, *MIMO Wireless Communication,* Cambridge University Press, 2010. ISBN:9780521137096.

[9] C. Cano and D.J. Leith, "Coexistence of Wi-Fi and LTE in unlicensed bands: A proportional fair allocation scheme," in *Proc. IEEE ICC workshop,* pp. 2288–2293, London, UK, June 8–12, 2015.

[10] F.S. Chaves, E. .L. Almeida, R.D. Vieira, A.M. Cavalcante, F.M. Abinader Jr., S. Choudhury, and K. Doppler, "LTE UL power control for the improvement of LTE/Wi-Fi coexistence," in *Proc. IEEE VT Fall,* 6 pages, Las Vegas, NV, Sept. 2–5, 2013.

[11] Q. Chen, G. Yu, H. Shan, A. Maaref, G.Y. Li, and A. Huang, "An opportunistic unlicensed spectrum utilization method for LTE and WiFi coexistence system," in *Proc. IEEE Globecom,* 6 pages, San Diego, CA, Dec. 6–12, 2015.

[12] Q. Chen, G. Yu, H. Shan, A. Maaref, G.Y. Li, and A. Huang, "Cellular meets WiFi: Traffic offloading or resource sharing?," *IEEE Transactions on Wireless Communications,* pp. 3354–3367, vol. 15, no. 5, Jan. 2016.

[13] L.-U. Choi and R.D. Murch, "A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach," *IEEE Trans. on Wireless Commun.,* vol. 3, no. 1, pp. 20–24, Jan. 2004.

[14] K. Chowdhury and I.F. Akyildiz, "CRP: A routing protocol for cognitive radio ad hoc networks," *IEEE Journal on Selected Areas in Communications,* vol. 29, no. 4, pp. 794–804, April 2011.

[15] M. Costa, "Writing on dirty paper," *IEEE Trans. Information Theory,* vol. 29, no. 3, pp. 439–441, May 1983.

[16] L. Ding, T. Melodia, S.N. Batalama, J.D.Matyjas, and M. Medley, "Cross-layer routing and dynamic spectrum allocation in cognitive radio ad hoc networks," *IEEE Trans. on Vehicular Technology,* vol. 59, no. 4, pp. 1969–1979, May 2010.

[17] D. Das, A.A. Abouzeid, and M. Codreanu, "Network layer scheduling and relaying in cooperative spectrum sharing networks," *IEEE Trans. on Wireless Communications,* April 2015.

[18] "Extending LTE advanced to unlicensed spectrum," Qualcomm whitepaper, Dec. 2013.

[19] P.K. Eswaran, A. Ravindran, and H. Moskowitz, "Algorithms for nonlinear integer bicriterion problems," *Journal of Optimization Theory and Applications,* vol. 63, no. 2, Nov. 1989.

[20] M. Ehrgott, *Multicriterion Optimization,* Springer-Verlag New York, 2005.

[21] S. Geirhofer, L. Tong, and B.M. Sadler, "Dynamic spectrum access in the time domain: Modeling and exploiting white space," *IEEE Communications Magazine,* vol. 45, no. 5, pp. 66–72, May 2007.

[22] A. Goldsmith, S.A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proceedings of the IEEE,* vol. 97, no. 5, pp. 894–914, May 2009.

[23] F. Gao, R. Zhang, Y.-C. Liang, and X. Wang, "Design of learning-based MIMO cognitive radio systems," *IEEE Transactions on Vehicular Technology,* vol. 59, no. 4, pp. 1707–1720, May 2010.

[24] Z. Guan and T. Melodia, "CU-LTE: Spectrally-efficient and fair coexistence between LTE and Wi-Fi in unlicensed bands," in *Proc. IEEE INFOCOM,* San Francisco, CA, USA, April 2016.

[25] A.T. Hoang, Y.C. Liang, and Y. Zeng "Adaptive joint scheduling of spectrum Sensing and data transmission in cognitive radio networks," *IEEE Trans. on Communications,* vol. 58, no. 1, pp. 235–246, Jan. 2010.

[26] Y.T. Hou, Y. Shi, and H.D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE Journal on Selected Areas in Commun.,* vol. 26, no. 1, pp. 146–155, Jan. 2008.

[27] Y.T. Hou, Y. Shi, and H.D. Sherali, *Applied Optimization Methods for Wireless Networks,* Cambridge University Press, 2014, ISBN-13: 978-1107018808.

[28] S. Hua, H. Liu, M. Wu, and S.S. Panwar, "Exploiting MIMO antennas in cooperative cognitive radio networks," in *Proc. IEEE INFOCOM,* pp. 2714–2722, Shanghai, China, April. 10-15, 2011.

[29] Y. Jian, C-F. Shih, B. Krishnaswamy, and R. Sivakumar, "Coexistence of Wi-Fi and LAA-LTE: Experimental evaluation, analysis and insights," in *IEEE International Conference on Communication Workshop (ICCW),* pp. 2325–2331, London, UK, June 8–12, 2015.

[30] S. Jafar and M. Fakhereddin, "Degrees of freedom for the MIMO interference channel," *IEEE Trans. on Information Theory,* vol. 53, no. 7, pp. 2637-2642, July 2007.

[31] S.K. Jayaweera, M. Bkassiny, and K.A. Avery, "Asymmetric cooperative communication based spectrum leasing via auctions in cognitive radio networks," *IEEE Trans. on Wireless Commun.,* vol. 10, no. 8, pp. 2716–2724, August 2011.

[32] D. Johnson, Y. Hu, D. Maltz, "The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4," IETF RFC 4728, Feb 2007.

[33] S.-J. Kim and G.B. Giannakis, "Optimal resource allocation for MIMO ad hoc cognitive radio networks," *IEEE Transactions on Information Theory,* vol. 57, no. 5, pp. 3117–3131, May 2011.

[34] J.N. Laneman, D.N.C. Tse, and G.W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. on Information Theory,* vol. 50, issue 12, pp. 3062–3080, Dec. 2004.

[35] H. Li and Z. Han, "Socially Optimal Queuing Control in Cognitive Radio Networks Subject to Service Interruptions: To Queue or Not to Queue?," *IEEE Trans. on Wireless Communications,* vol. 10, no. 5, May 2011.

[36] Y. Li, F. Baccelli, J.G. Andrews, T.D. Novlan and J. Zhang, "Modeling and analyzing the coexistence of licensed-assisted access LTE and Wi-Fi," in *2015 IEEE Globecom Workshops (GC Wkshps),* 6 pages, San Diego, CA, Dec. 6–10, 2015.

[37] J. Liu, Y.T. Hou, H.D. Sherali, "Routing and power allocation optimization for MIMO-based ad hoc networks with dirty paper coding," in *Proc. IEEE International Conference on Communications (ICC) – Wireless Networking Symposium,* pp. 2859–2864 May 19–23, 2008, Beijing, China.

[38] F. Liu, E. Bala, E. Erkip, and R. Yang, "A framework for femtocells to access both licensed and unlicensed bands," in *Proc. WiOpt,* pp. 407–411, Princeton, NJ, May 9–13, 2011.

[39] N.A. Lynch, *Distributed Algorithms,* Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1996. ISBN: 1558603484.

[40] "LTE license assisted access," Ericsson whitepaper, Jan. 2015.

[41] Zh. Ma, W. Chen, K.b. Lataief, and Z. Cao, "A Semi Range-Based Iterative Localization Algorithm for Cognitive Radio Networks," *IEEE Transaction on Vehicular Technology,* vol. 59, no. 2, pp. 704-717, Feb. 2010.

[42] R. Manna, R.H.Y. Louie, Y. Li, and B. Vucetic, "Cooperative spectrum sharing in cognitive radio networks with multiple antennas," *IEEE Trans. on Signal Processing,* vol. 59, no. 11, pp. 5509–5522, Nov. 2011.

[43] T. Nadkar, V. Thumar, G. Shenoy, A. Mehta, U.B. Desai, and S.N. Mechant, "A cross-layer framework for symbiotic relaying in cognitive radio networks," in *Proc. IEEE DySPAN,* pp. 498–509, Aachen, Germany, May. 3–6, 2011.

[44] T. Nihtila, V. Tykhomtro, O. Alanen, M.A. Uusitalo, A. Sorri, M. Moisio, S. Iraji, R. Ratasuk, and N. Mangalvedhe, "System performance of LTE and IEEE 802.11 coexisting on a shared frequency band," in *Proc. IEEE WCNC,* pp. 1038–1043, Shanghai, China, April 2013.

[45] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc on-demand distance vector (AODV) routing," IETF RFC 3561, July 2003.

[46] President's Council of Advisors on Science and Technology (PCAST), "Report to the President — Realizing the Full Potential of Government-held Spectrum to Spur Economic Growth," July 2012, available online http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast_spectrum_report_final_july_20_2012.pdf.

[47] "Qualcomm research LTE in unlicensed spectrum: harmonious coexistence with Wi-Fi," Qualcomm whitepaper, June 2014.

[48] B. Radunovic and J.-Y. Le Boudec, "Rate performance objectives of multi-hop wireless networks," in *Proc. IEEE INFOCOM,* pp. 1916–1927, Hong Kong, China, March 7–11, 2004.

[49] H. Rahul, S. Kumar, and D. Katabi, "JMB: scaling wireless capacity with user demands," in *Proc. ACM SIGCOMM,* pp. 235–246, Helsinki, Finland, Aug. 2012.

[50] T. K. Ralphs, M. J. Saltzman, and M.M. Wiecek, "An improved algorithm for biobjective integer programming," *Annals of Operations Research,* vol. 147, pp. 43–70, 2006.

[51] R. Ratasuk, M.A. Uusitalo, N. Mangalvedhe, A. Sorri, S. Iraji, C. Wijting, and A. Ghosh, "License-exempt LTE deployment in heterogeneous network," in *Proc. IEEE ISWCS,* pp. 246–250, Paris, France, Aug. 28–31, 2012.

[52] S. Rosloniec, *Fundamental Numerical Methods for Electrical Engineering,* Springer, Berlin, 2008.

[53] S. Sagari, S. Baysting, D. Saha, I. Seskar, W. Trappe, and D. Raychaudhuri, "Coordinated dynamic spectrum management of LTE-U and Wi-Fi networks," in *Proc. IEEE DySPAN,* pp. 209–222, Stockholm, Sweden, Sept. 29–Oct. 2, 2015.

[54] S.Sagari, "Coexistence of LTE and Wi-Fi heterogeneous networks via inter network cooperation," in *ACM MobiSys PhD Forum,* 2 pages, Bretton Woods, NH, USA, June 16–19, 2014.

[55] S. Sagari, I. Seskar, and D. Raychaudhuri, "Modeling the coexistence of LTE and WiFi heterogeneous networks in dense deployment scenarios," in *Proc. IEEE ICC workshop,* pp. 2301–2306, London, UK, June 8–12, 2015.

[56] W. Su, J.D. Natyjas, and S. Batalama, "Active cooperation between primary users and cognitive radio users in cognitive ad-hoc network" in *Proc. IEEE ICASSP,* pp. 3174–3177, March. 14–19, 2010.

[57] A. Schrijver, *Theory of Linear and Integer Programming,* Wiley Interscience, New York, NY, 1986.

[58] S. Sengupta and K.P. Subbalakshmi, "Open research issues in multi-hop cognitive radio networks," *IEEE Communication Magazine,* vol. 52, no. 4, pp. 168–176, April, 2013.

[59] Y. Shi, J. Liu, C. Jiang, C. Gao, and Y.T. Hou, "A DoF-based link layer model for multi-hop mino networks," *IEEE Trans. on Mobile Computing,* vol. 12, issue 7, pp. 1395–1408, 2014.

[60] Y. Shi, Y.T. Hou, J. Liu, and S. Kompella, "Bridging the gap between protocol and physical models for wireless networks," *IEEE Trans. on Mobile Computing,* vol. 12, issue 7, pp. 1404–1416, July 2013.

[61] O. Simone, I. Stanojev, S. Savazzi, Y. Bar-Ness, U. Spagnolini, and R. Pickholtz, "Spectrum leasing to cooperating secondary ad hoc networks," *IEEE Journal on Selected Areas in Commun.,* vol. 26, no. 1, pp. 203–213, Jan. 2008.

[62] G.S. Smith, "A Direct Derivation of a Single-Antenna Reciprocity Relation for the Time Domain." in *IEEE Trans. on Antennas and Propagation,* vol. 52, no. 6, pp. 1568–1577, June 2004.

[63] G.S. Smith, "A direct derivation of a single-antenna reciprocity relation for the time domain," *IEEE Trans. on Antennas and Propagation,* vol. 52, no. 6, pp. 1568–1577, June 2004.

[64] Y. Song, K.W. Sung, and Y. Han, "Coexistence of Wi-Fi and cellular with listen-before-talk in unlicensed spectrum," *IEEE Communications Letters,* pp. 161–164, Dec. 2015.

[65] Q.H. Spencer, A.L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. on Signal Processing,* vol. 52, no. 2, pp. 461–471, Feb. 2004.

[66] R.E. Steuer and E. Choo, "An interactive weighted tchebycheff procedure for multiple objective programming," *Mathematical Programming,* vol. 26, issue 3, pp. 326–344, 1983.

[67] K. Tang, "LTE on unlicensed spectrum: innovation and coexistence," August 18, 2015, http://www.wca.org/wp-content/uploads/2015/11/ Qualcomm_LTEU_LAA_WCA _08182015.pdf, accessed April 22, 2016.

[68] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication,* Chapter 7, Cambridge University Press, Cambridge, UK, 2005.

[69] "U-LTE:Unlicensed spectrum utilization of LTE," Huawei Technologies, 2014.

[70] R. Urgaonkar and M.J. Neely, "Opportunistic cooperation in cognitive femtocell Networks," *IEEE Journal on Selected Areas in Communications,* vol. 30, no. 3, pp. 607–616, April 2012.

[71] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Information Theory,* vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[72] A.M. Wyglinski, M. Nekovee, and Y.T. Hou, *Cognitive Radio Communications and Networks: Principles and Practice.* Academic Press/Elsevier, 2010. ISBN-13:9780123747150.

[73] J. Wang, J. Chen, D. Cabric, "Cramer-Rao Bounds for Joint RSS/DoA-Based Primary-User Localization in Cognitive Radio Networks," *IEEE Transaction on Wireless Communications,* vol. 12, no. 3, pp. 1363–1375, Mar. 2013.

[74] X. Xie, X. Zhang, and K. Sundaresan, "Adaptive feedback compression for MIMO networks," in *Proc. of ACM MobiCom,* pp. 477–488, Miami, FL, Sep. 2013.

[75] X. Yuan, C. Jiang, Y. Shi, Y.T. Hou, W. Lou, S. Kompella, and S.F. Midkiff, "Toward transparent coexistence for multi-hop secondary cognitive radio networks," *IEEE Journal on Selected Areas in Communications,* vol. 33, no. 5, pp. 958–971, May 2015.

[76] X. Yuan, C. Jiang, Y. Shi, Y.T. Hou, W. Lou, and S. Kompella, "Beyond interference avoidance: on transparent coexistence for multi-hop secondary CR networks," in *Proc. IEEE SECON,* pp. 398–405, New Orleans, LA, June 24–27, 2013.

[77] X. Yuan, Y. Shi, Y.T. Hou, W. Lou, S.F. Midkiff, and S. Kompella, "Achieving Transparent Coexistence in a Multi-hop Secondary Network Through Distributed Computation," in *Proc. IEEE IPCCC,* Austin, TX, Dec. 5–7, 2014.

[78] X. Yuan, X. Qin, Y. Shi, Y.T. Hou, W. Lou, S.F. Midkiff, and S. Kompella, "A Distributed Algorithm to Achieve Transparent Coexistence for a Secondary Multi-hop MIMO Network," The Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, July 2015. URL: https://docs.google.com/a/vt.edu/viewer?a=v&pid=sites&srcid=dnQuZWR1fHh1MXxne DozY2ViNzljNzg0OGEzZjg2.

[79] X. Yuan, Y. Shi, Y.T. Hou, W. Lou, and S. Kompella, "UPS: A United Cooperative Paradigm for Primary and Secondary Networks," in *Proc. IEEE MASS,* Hangzhou, China, Oct. 14–16, 2013.

[80] S. Yun and L. Qiu, "Supporting Wi-Fi and LTE coexistence," in *Proc. IEEE INFOCOM,* pp. 810–818, Hong Kong, April 26–May 1, 2015.

[81] S. Zaks, "Optimal distributed algorithms for sorting and ranking," *IEEE Trans. on Computers,* vol. 5, no. 1, pp. 376–379, April 1985.

[82] D. Zhang, Z. Tian, and G. Wei, "Spatial capacity of narrowband vs. ultra-wideband Cognitive Radio Systems," *IEEE Trans. on Wireless Communications,* vol. 7, no. 11, pp. 4670–4680, Nov. 2008.

[83] J. Zhang and Q. Zhang, "Stackelberg game for utility-based cooperative radio network," in *Proc. ACM MobiHoc,* pp. 23–32, New Orleans, LA, USA, May 18–21, 2009.

[84] X. Zhang, K. Sundaresan, M.A. Khojastepour, S. Rangarajan, and K.G. Shin, "NEMOx: scalable network MIMO for wireless networks," in *Proc. ACM MobiCom,* pp. 453-464, Miami, FL, Sep. 2013.

[85] R. Zhang and Y.-C. Liang, "Exploiting multi-antennas for opportunistic spectrum sharing in cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing,* vol. 2, no. 1, pp. 88–102, February 2008.

[86] Y.J. Zhang and A.M.-C. So, "Optimal spectrum sharing in MIMO cognitive radio networks via semidefinite programming," *IEEE Journal on Selected Areas in Commun.,* vol. 29, no. 2, pp. 362–373, February 2011.

[87] R. Zhang, M. Wang, L.X. Cai, X.Shen, L.L. Xie, and Y. Cheng, "Modeling and analysis of MAC protocol for LTE-U coexisting with Wi-Fi," in *Proc. IEEE GLOBECOM,* 6 pages, San Diego, CA, Dec. 6–10, 2015.

[88] H. Zhang, X. Chu, W. Guo, and S. Wang, "Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum," *IEEE Communications Magazine,* vo. 53, no. 3, pp. 158–164, March 2015.

[89] Q. Zhao and B.M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing Magazine,* vol. 24, no. 3, pp. 79–89, May 2007.

[90] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. on Information Theory,* vol. 49, no. 5, pp. 1073–1096, May 2003.