BIG DATA PRIVACY PRESERVATION FOR CYBER-PHYSICAL SYSTEMS

A Dissertation

Presented to

the Faculty of the Electrical and Computer Engineering

University of Houston

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in Electrical and Computer Engineering

by

Jingyi Wang

May 2019

BIG DATA PRIVACY PRESERVATION FOR CYBER-PHYSICAL SYSTEMS

Jingyi Wang

Approved:

Chair of the Committee
Dr. Miao Pan, Assistant Professor,
Electrical and Computer Engineering

Committee Members:

Dr. Zhu Han, Professor,
Computer Science
Electrical and Computer Engineering

Dr. Saurabh Prasad, Assistant Professor,
Electrical and Computer Engineering

Dr. Rose Faghih, Assistant Professor,
Electrical and Computer Engineering

Dr. Lijun Qian, Professor,
Electrical and Computer Engineering,
Prairie View A&M University

Dr. Suresh K. Khator, Associate Dean,
Cullen College of Engineering

Dr. Badrinath Roysam, Professor and Chair,
Electrical and Computer Engineering

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Miao Pan for his general guidance and continuous support during my graduate studies. His keen insight, profound academic background and optimistic personality cast light upon the mist of my studies and pointed the way towards a bright future. Whatever in life and in research, I learned a lot from his words and deeds, which are precious assets for me in my lifetime.

Furthermore, I would like to express my sincere thanks to my committee member, Dr. Zhu Han, Dr. Saurabh Prasad, Dr. Rose Faghih and Dr. Lijun Qian, who provided continuous encouragement and valuable advice in my research. I would also like to thank all other scholars I have collaborated with, Dr. Wenbo Ding, Dr. Jian Song, Dr. Yuanxiong Guo, Dr. Hongyan Li, Dr. Zaixin Lu, Dr. Yanmin Gong, Dr. Minglei Shu, Dr. Yinglong Wang, Dr. Riku Jantii, Dr. Qixun Zhang and Dr. Zhiyong Feng. Many great ideas came out through discussions with them, and I really appreciate their helpful comments and suggestions.

My thankfulness also goes to my dear friends and colleagues, Dr. Sai Mounika Errapotu, Xinyue Zhang, Debing Wei, Jiahao Ding, Dian Shi, Pavana Prakash, Rui Chen, Steban Soto, Liang Li, Dr. Jixiang Lu, Huaqing Zhang, Xinyan Li, Joshua Jones and many others for creating such an enjoyable studying environment. My life became much more colorful with the company of all of them.

Last but by no means least, I owe my gratitude towards my parents and my family, who provided priceless love and sustainable support throughout my studies and my life. Thanks for everything that you give me.

BIG DATA PRIVACY PRESERVATION FOR CYBER-PHYSICAL SYSTEMS

An Abstract

of a

Dissertation

Presented to

the Faculty of the Electrical and Computer Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in Electrical and Computer Engineering

by

Jingyi Wang

May 2019

Abstract

Cyber-physical systems (CPS) often referred as “next generation of engineered systems" are sensing and communication systems that offer tight integration of computation and networking capabilities to monitor and control entities in the physical world. The advent of cloud computing technologies, artificial intelligence and machine learning models has extensively contributed to these multidimensional and complex systems by facilitating a systematic transformation of massive data into information. Though CPS have infiltrated into many areas due to their advantages, big data analytics and privacy are major considerations for building efficient and high-confidence CPS. Many domains of CPS such as smart metering, intelligent transportation, health care, sensor/data aggregation, crowd sensing etc., typically collect huge amounts of data for decision making, where the data may include individual or sensitive information. Since vast amount of information is analyzed, released and calculated by the system to make smart decisions, big data plays a key role as an advanced analysis technique providing more efficient and complete solutions for CPS. However, data privacy breaches during any stage of these large scale systems, either during collection or big data analysis can be an undesirable loss of privacy for the participants and for the entire system.

This work focuses on effective big data analytics for CPS and addresses the privacy issues that arise in various CPS applications. Because of their numerous advantages, CPS and its communication networks inevitably become the targets of attackers and malicious users either during data collection, data storage, data transmission, or data processing and computation, keeping users’ information at risk. Given these challenges, this work endeavors to develop a series of privacy preserving data analytic and processing methodologies through data driven optimization based on differential privacy; and focuses on effectively integrating the data analysis and data privacy preservation techniques to provide the most desirable solutions for the state-of-the-art CPS with various application-specific requirements.

Table of Contents

Acknowledgements **v**

Abstract **vii**

Table of Contents **viii**

List of Figures **xii**

List of Tables **xiv**

1 Introduction **1**

2 Differential Privacy and Big Data **6**

2.1 Preliminaries 6

2.1.1 Centralized Differential Privacy 6

2.1.2 Distributed Differential Privacy 8

2.1.3 Local Differential Privacy 9

2.2 Big Data Analysis : Data-driven Methodology Preliminaries 11

2.2.1 ζ-structure metric 11

3 Optimization Based Primary Users' Operational Privacy Preservation **15**

3.1 Introduction 15

3.1.1 Related Work 15

3.1.2 Our Contribution 18

3.2 System Description 19

3.2.1 Network Configuration 19
3.2.2 Other Related Model in the System 20
3.2.3 Attack Model 22
3.3 Obfuscation Strategy and Problem Formulation 23
3.3.1 Utility Functions of PUs and SUs 23
3.3.2 PUs' Operational Privacy Preserving Optimization 25
3.4 Risk-Averse Stochastic Programming for Preserving Temporal Operational Privacy 29
3.4.1 Reference Distribution 29
3.4.2 Converge Rate under ζ-Probability Metrics 29
3.5 Performance Evaluation 33
3.6 Conclusion 36

4 Spectrum Trading with Secondary Users' Differential Privacy Preservation 39

4.1 Introduction 39
4.2 System Description and 3DPP Outline 43
4.2.1 System Model and Adversary Model 43
4.2.2 3DPP Outline 44
4.3 3DPP Problem Formulation 46
4.3.1 PSP's Revenue Maximization Formulation 46
4.3.2 Data-Driven Based PSP's Revenue Optimization 47
4.3.3 3DPP: Data-Driven Based PSP's Revenue Optimization under ϵ-DP . 48
4.4 3DPP Proof and Solutions 49

4.4.1 Problem Reformulation under ζ-Probability Metrics, and Solutions . . 53
4.5 Performance Evaluation . . . 54
4.5.1 Simulation Setup . . . 54
4.5.2 Privacy and Performance Analysis . . . 55
4.6 Related Work . . . 58
4.7 Conclusion . . . 59

5 Optimization for Utility Providers with Differential Privacy of Users' Energy Profile 61

5.1 Introduction . . . 61
5.2 Network Model . . . 64
5.2.1 Add Noise with Distributed Differential Privacy Algorithm . . . 65
5.2.2 Data-driven Prediction . . . 66
5.2.3 Cost Minimization Problem Formulation . . . 68
5.2.4 Solution to the Optimization Problem . . . 70
5.3 Performance Evaluation . . . 74
5.4 Conclusion . . . 74

6 Caching with Users' Local Differential Privacy in Information-Centric Networks 76

6.1 Introduction . . . 76
6.2 Network Model and Preliminaries . . . 79
6.2.1 System Description . . . 79
6.3 Data-Driven Caching Revenue Maximization Problem with Local Differential Privacy . . . 80

6.3.1 Protecting Private Content Preference with Local Differential Privacy 80

6.3.2 Data-driven Analysis of Content Popularity 81

6.3.3 Caching Revenue Maximization Problem with Local Privacy Preservation 82

6.3.4 Solution to Caching Optimization under Distribution Uncertainty . . . 83

6.4 Performance Evaluation 85

6.5 Conclusion 86

7 Data-Driven Small Cell Placement Optimization with Users' Differential Privacy for Wireless NGNs 88

7.1 Introduction 89

7.1.1 Related Work 90

7.1.2 Our Contribution 90

7.2 System Description 93

7.2.1 Network Configuration 93

7.2.2 Revenue Maximization Problem Formulation 94

7.3 Solution to The Optimization Problem 98

7.4 Performance Evaluation 104

7.4.1 Simulation Setup 104

7.4.2 Privacy and Performance Analysis 104

7.5 Conclusion 105

8 Future Works 106

References 109

List of Figures

1.1 Cyber-Physical Systems 2

2.1 Wasserstein metric (one-dimensional case). 12

2.2 uniform metric 12

3.1 System architecture and temporal operational attacks. 17

3.2 A toy overall conflict graph observed by a PU. 19

3.3 Impact of size of historical data on system utility (One SU). 30

3.4 Impact of size of historical data on system utility (One SU) under different coefficient c=1, b=4. 32

3.5 Impact of historical data on system utility (One SU) under different distribution. 33

3.6 Impact of historical data on system utility (10 SUs). 34

3.7 Impact of size of historical data and different number of SUs under Uniform metric. 35

3.8 Impact of different number of SUs on system utility under different metrics. . 36

3.9 PU temporal operational privacy and system utility tradeoff, $|\mathcal{N}| = 1$. 37

3.10 Temporal operational privacy and system utility tradeoff, $|\mathcal{N}| = 3$ and $|\mathcal{N}| = 5$. 37

4.1 Illustrative examples for the traffic demand privacy breach of SUs in spectrum trading. 40

4.2 The spectrum trading procedure of 3DPP. 43

4.3 Data-Driven spectrum trading without ϵ-DP. 52

4.4 Total revenue of PSP under different probability distance metrics. 52

4.5 Total revenue of the PSP with 3DPP under Fortet-Mourier metric 55

4.6 Total revenue of the PSP with 3DPP under Kantorovich metric metric 55

4.7 Total revenue of the PSP with 3DPP under Uniform metric metric 56

4.8 Total revenue of the PSP with 3DPP under different confidence levels. 56

5.1 Network overview. 65

5.2 Distribution of consumers' energy demand under different ϵ. 73

5.3 Total cost under L_1 norm 73

5.4 Total cost under L_∞ norm 75

6.1 System description. 80

6.2 Expected revenue without users' privacy preservation. 83

6.3 Performance under Kantorovich metric 86

6.4 Performance under Fortet-Mourier metric. 86

6.5 Performance under Uniform metric. 87

7.1 Illustration for small cell deployment in NGNs. 91

7.2 Users' per hour transmission data distribution 101

7.3 Total revenue under different metrics without different privacy. 102

7.4 Total revenue under different metrics with privacy budget ϵ=0.3. 102

7.5 Total revenue under Kantorovich metric. 102

7.6 Total revenue under Fortet-Mourier metric. 103

7.7 Total revenue under Uniform metric without different privacy. 103

List of Tables

3.1 The list of notations 28

4.1 The list of notations 45

Chapter 1

Introduction

A cyber-physical system (CPS) is a sensing and communication system that offers tight integration and combination of computation, networking and physical processes. CPS are largely referred to as the next generation of engineered systems with the integration of communication, computation, and control to achieve the goals of stability, performance, robustness, and efficiency for physical systems. CPS mainly consists of two components: a physical system and a cyber system. Typically, embedded computers and sensors in the physical system collect measurements that reflect the current state of the physical system, and then send them to the cyber system through communication networks in real time. At the same time, the cyber system processes the received measurements and obtains the status of the physical system. Based on the processing results, the cyber system responds to the physical system in real time by sending directive instructions through communication networks, achieving better performance, or maintaining system stability.

CPS have infiltrated into many areas, such as aerospace, automobiles, chemical processing, civil infrastructure, energy, healthcare, manufacturing, transportation, entertainment, and consumer appliances. Big data analysis is a major issues for building efficient cyber-physical systems. The advent of the data processing and big data analysis has contributed significantly to the growth of cyber-physical systems. Big data shows great potential in decision making, optimizing operations and capitalizing on new sources of revenues in a variety of fields. From the enreach data, the CPS can organize the prognostic and health management, detect invisible problem and avoid unplanned downtimes to overwhelm uncertainties in the system. Analysing and studying the data efficiently from the cyber-physical system and is a promising way to remaining useful life prediction, fault diagnose and fault detection.

Privacy is another consideration in CPS. Cyber-physical systems are often distributed

Figure 1.1: Cyber-Physical Systems

across wide geographic areas and typically collect huge amounts of information for data analysis and decision making. For example, the operation of emerging large-scale monitoring and control systems, such as intelligent transportation systems or smart grids relies on information continuously provided by and about their users. The collection of information helps the system make smart decisions through sophisticated machine learning algorithms. But, it can be an undesirable loss of privacy for the participants, thereby putting their promised benefits at risk. Data breaches, however, could potentially happen in any part of the system, including the stages of data collection, data transmission, data operation, and data storage. Due to its importance, CPS and its communication networks inevitably become the targets of attackers and malicious users. More recently, the need has arisen for new theories and tools that can protect the individuals around whom sensor networks and other smart information sources are being built for purposes of collecting dynamic data.

Due to emerging computing technologies, CPS has caught much attention in the research community, especially in the areas of security and privacy to prevent it from outside attackers and malicious users. This work focuses on developing tools and techniques for effectively preserving the privacy of the users in various CPS. Since cyber-physical systems have

applications in various domains, the traditional ways of protecting privacy do not readily apply to these emerging distributed systems. Considering application-specifi c requirements for different CPS, our inter-disciplinary research integrates various mathematical models that control the physical system along with data-driven methodology and differential privacy based techniques that deal with the uncertainty of the CPS simultaneously. The major goal of this work is to maximize the utility of the system with limited uncertain sampled data and guarantee the privacy of individuals or corporations for real-time practical CPS applications. The notion of privacy and uncertainty for CPS is considered as follows.

Privacy: CPS and its communication networks usually involve in many entities, including human beings and companies, and sensors continuously collect data about them. CPS privacy mainly consists of two parts. The first is identity privacy, such as whether an entity participates in a CPS or communication network. The other is data privacy. The data collected by sensors or the data collected during analysis may reveal sensitive information about human beings and corporations. For instance, smart meters in a power system can be used to infer household activities. This work focuses on privacy preservation during data collection, transmission and computation.

Uncertainty: CPS and its communication networks usually involve in many entities, including human beings and companies, and sensors continuously collect data about them. Sometimes, the size of the data maybe too large to process. For instance, the Los Angeles Department of Water and Power(LADWP) serves 4.1 million consumers, and has a net generation capacity of 7,100MW. It is very hard if not impossible to process all of the data in the realtime. Therefore, in our work, we collect the data from the sampled users to learn the characteristic of the whole data set. How to deal with the uncertainty of the data is very challenge. For instance, when the system collect the data from sampled users, the result cannot be 100% accurate to represent the whole user's characteristic. There always will be a gap between the sampled result and real result. In our work, we mathematically describe this gap and consider it in the optimization problem to maximize the utility of the CPS

under the real scenario.

This dissertation focuses on the privacy preservation and data-driven analysis in CPS, including its communication networks. As we can see, CPS is a large area, and it is challenging to cover all aspects of this area. Different from the existing works, we try to address the privacy issues in real world scenarios and guarantee the utility of the underlying system without compromising the privacy and considering the uncertainty of the unknown whole dataset. Specifically, we present three different CPS applications, considering their privacy requirements and the mathematically distribution distance between sample dataset and whole dataset, and we present our works on guaranteeing privacy preservation in these applications. In this chapter, we will briefly state each problem, and describe the contribution of each work.

Specifically, we first illustrate the differential privacy technique and the data-driven methodology we employed in big data analysis in Chapter2. We present an obfuscation strategy for PUs in spectrum trading to preserve PUs' temporal operational privacy, which employ a data-driven risk-averse model to characterize the uncertainty of SUs' demands and jointly consider the frequency reuse in the cognitive radio (CR) network in Chapter 3. Then, we present a data-driven spectrum trading scheme which maximizes PUs' revenue. Meanwhile, integrate centralized differential privacy in the spectrum trading to protect SUs' demand privacy. In Chapter 5, we introduce potential privacy attacks in one of the most important cyber system, i.e., smart grids with focus on privacy issues in consumers, and propose efficient countermeasures to defend against such attacks. In Chapter 6, we present a novel scheme in information-centric network to predict the content popularity from limited sample users, and offer the cache-enabled access points (APs) to enjoy the benefits of caching users' preferable contents without disclosing the users' privacy. In Chapter 7, we integrate differential privacy (DP) preserving techniques into data-driven optimization, and propose a novel scheme that not only preserves the privacy of 5G next generation networks users' transmission information, but also maximizes the revenue of small cell deployment. At

last, we present some possible future works to which data-driven methodology and privacy technique can be applied in Chapter 8.

Chapter 2

Differential Privacy and Big Data

2.1 Preliminaries

2.1.1 Centralized Differential Privacy

Differential privacy ensures that any sequence of output from data set (e.g., responses to queries) is "essentially" equally likely to occur, no matter any individual item is present or not presenst [1–3]. In other word, a single item in the database does not (significantly) affect the outcome of analysis. The differential privacy technique keeps the characteristic of the whole data set, and preserves privacy of each individual data item.

In the standard (or centralized) DP setting, each user sends raw data to the database, who obtains the true distribution, adds noise, and then publishes the result. In this setting, the aggregator is trusted to not reveal the raw data and is trusted to handle the raw data correctly [4]. DP keeps the characteristic of the whole data set, and preserves information privacy of each individual.

We assume a database x being collection of records from a universe $\mathcal{X}$[1]. Each entry d_i represents the number of elements in the database x of *type* $i \in \mathcal{X}$, and $\mathbb{N}$ denotes the set of all non-negative integers, including zero. It will be convenient to represent database by their histograms: $x \in \mathbb{N}^{|\mathcal{X}|}$. In this presentation, the l_1 distance between two database d and d' is defined as follows [1].

Definition 2.1. Distance Between Database*: The l_1 norm of a database d is denoted as $||x||_1$ and is defined to be:*

$$||x||_1 = \sum_{i=1}^{|\mathcal{X}|} |x_i|. \tag{2.1}$$

[1]A universe $\mathcal{X}$ of data type indicates the set of all possible database rows.

The l_1 distance between two database d and d' (i.e., $||x - x'||_1$) is a measure of how many records differ between x and x'.

Definition 2.2. Differential Privacy*: Let $\mathcal{A}$ denote a randomized algorithm, $\boldsymbol{r}$ denote the output result and x denote the input (data set), i.e., $\mathcal{A}(x) = \boldsymbol{r}$. For all $x, x' \subseteq \mathbb{N}^{|\mathcal{X}|}$ satisfies $||x - x'||_1 \leq 1$,*

$$\log \frac{\Pr(\boldsymbol{r}|x)}{\Pr(\boldsymbol{r}|x')} \leq \epsilon. \tag{2.2}$$

Then we say $\mathcal{A}$ satisfies ϵ-DP (differential privacy). The parameter ϵ represents the privacy budget the algorithm $\mathcal{A}$ offered. A smaller value of ϵ indicates the stronger privacy level guarantee and more perturbation noise; a larger value of ϵ means a weaker privacy level guarantee with higher data utility.

Numeric queries, function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ map databases to k real numbers. The l_1 sensitivity determines how accurately we can answer such queries.

Definition 2.3. Function f with l_1-sensitivity*: The l_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ is*

$$\Delta f = \max_{\substack{d,d' \in \mathbb{N}^{|\mathcal{X}|}, \\ ||d-d'||_1 = 1}} ||f(x) - f(x')||_1. \tag{2.3}$$

The l_1 sensitivity of a function f captures the magnitude, by which an individual item can change the function f in the worst case. To understand it intuitively, f could be the query to the database, and $f(x)$ is the answer of f under the database x. The sensitivity Δf in the response is the key element to hide the information of a single individual [1].

Definition 2.4. The Laplace Distribution and the Laplace Mechanism*: The PDF (Probability Density of Function) of the Laplace Distribution (centered at 0) with scale b is:*

$$\text{Lap}(z|b) = \frac{1}{2b} \exp\Big(- \frac{|z|}{b} \Big). \tag{2.4}$$

In the following work, we will write $\text{Lap}(b)$ as the simplification to a random variable

$Z \sim \text{Lap}(b)$. The *Laplace Mechanism* simply computes f, and perturb each coordinate with noise drawn from the Laplace distribution, where $b = \Delta f/\epsilon$. The scale of the noise is calibrated to the sensitivity of f as described in Definition 3. The Laplace Mechanism $\mathcal{A}_L$ is defined as

$$\mathcal{A}_L(d, f(\cdot), \epsilon) = f(d) + (Y_1, \cdots, Y_k),$$

where Y_i are i.i.d random variables drawn from $\text{Lap}(\Delta f/\epsilon)$. The proof of Laplace mechanism reserves ϵ-DP is shown in [1].

2.1.2 Distributed Differential Privacy

As introduced in the previous section, the centralized differential privacy setting has a strong assumption that a trustworthy third-party database or data aggregator is required to apply the randomized algorithm on the exact data of data providers. In reality, people may evade to provide data to a survey including sensitive questions. In such a situation that the data providers trust no one even the data collectors but only themselves, secure multi-party computation and homomorphic encryption are suitable to involve in the differential privacy definition. In [5], the authors propose a private stream aggregation algorithm which guarantees distributed differential privacy of individual user, while the aggregator can only get the statistical results, but not learn any other unintended information from users. We assume $\mathbf{x} = \{x_1, \cdots, x_n\}$ denotes the vector of users' data, function f represents the desired statistics of data provider and $\mathcal{O}$ indicates the output range of function f. Then, the definition of *distributed differential privacy* (DDP) is shown as follows.

Definition 2.5. *With a privacy confidence parameter $\epsilon > 0$ and $0 \leq \delta < 1$, a randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-distributed differential privacy with respect to the function f, for any subset $X \subseteq \mathcal{O}$, when given two neighbor vectors $\mathbf{x}, \mathbf{x}'$ [5]:*

$$\Pr[f(\mathcal{A}(\mathbf{x})) \in X] \leq e^{\epsilon} \cdot \Pr[f(\mathcal{A}(\mathbf{x}')) \in X] + \delta.$$

The two neighbor vectors $\mathbf{x}, \mathbf{x}'$ are supposed to have only one element different. The privacy confidence parameter ϵ controls the privacy preservation level. With smaller ϵ, it is less possible to distinguish the outputs of the randomized algorithm $\mathcal{A}$ with two different inputs, which means the privacy protection is stronger.

In the setup of the DDP algorithm, based on the homomorphic encryption scheme, each customer gets a private key sk_j to encrypt the data with distributed differential private noise and the data provider also gets the key sk_0 to decrypt the statistics when receiving all of the cipher texts from customers. Because the data provider can only learn the noisy statistic, each user would add less noise to the demand data, if the randomization of the desired statistic $f(\mathcal{A}(\mathbf{x}))$ is big enough. As the user's data is in a discrete group, in DDP, a symmetric geometric distribution is exploited to guarantee the satisfaction of differential privacy. The probability mass function of symmetric geometric distribution is $\text{Geom}(\alpha) = \frac{\alpha-1}{\alpha+1}\alpha^{-|j|}$. When α is set to $e^{\epsilon/\Delta}$ and the noise comes from this symmetric geometric distribution, the randomized algorithm $\mathcal{A}$ is able to achieve differential privacy [5].

2.1.3 Local Differential Privacy

Differential privacy [6] is used to obtain the statistical information of databases without disclosure of the data providers' privacy. Intuitively, given two databases, which have only one element different from each other, as the inputs of a randomization algorithm, the outputs are not distinguishable. However, there must exist a trustworthy database or data aggregator when applying the centralized differential privacy. In local differential privacy, it assumes that the service database is *honest-but-curious*, which means the privacy leakage possibility increases. Therefore, local privacy setting is suitable in the situation that the data providers trust no one except themselves. The Warner's random response model [7] is one of the oldest local privacy model applied in survey sampling. If there are two answers of one question, the data provider will reply truly with probability of p and falsely with probability of $1 - p$. Combining local privacy and differential privacy, the definition of *local*

differential privacy (LDP) is shown as follows.

Definition 2.6. *With a privacy confidence parameter $\epsilon \geq 0$, a randomized algorithm $\mathcal{A}$ satisfies ϵ-local differential privacy, when given two inputs x and x' [8]:*

$$\frac{\Pr[\mathcal{A}(x) = z]}{\Pr[\mathcal{A}(x') = z]} \leq e^{\epsilon},$$

where z is the secure view of the input.

Therefore, with a specific output s from the randomized algorithm $\mathcal{A}$, it is not able to determine or can infer with negligible probability whether the input is x or x'. Additionally, the privacy confidence parameter ϵ controls the privacy preservation level, which means there is more possibility to distinguish the outputs of the randomized algorithm $\mathcal{A}$ with two different inputs with a higher value of ϵ. In other words, smaller ϵ means higher privacy preservation level.

To perform a LDP mechanism, it contains several steps. First, the true data is encoded locally into a vector or a number. Next, the encoded data is randomized by a specific function. At last, the processed data will be sent to the data aggregator or database. Among the three steps, the combination of the first two steps is the randomized algorithm $\mathcal{A}$ in the definition, which is finished locally and supposed to satisfy ϵ-LDP. In [9], the authors have introduced an optimized LDP protocol, named optimal local hashing (OLH), which can offer higher accuracy of frequency estimation with lower communication cost.

In the encoding step of OLH, the input, denoted as $r_u \in [1, F]$, is first encoded with the hash function H, which can hash the input value into $[g]$ ($g > 2$), uniformly chosen from a universe hash function family $\mathbb{H}$. The output of first step is represented $r_u^H = Encode(r_u) = \langle H, r_u \rangle$. In the next perturbation step, r_u^H is perturbed into $r'_u = Perturb(r_u^H)$, with the probability shown as

$$\forall_{k\in[g]}\Pr[r'_u = k] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon+g-1}, \text{when } r_u^H = k, \\ \\ q = \frac{1}{e^\epsilon+g-1}, \text{when } r_u^H \neq k. \end{cases} \tag{2.5}$$

This local hashing protocol including encoding and perturbation is satisfactory to ϵ-LDP (the detailed proof shown in [9]). With $g = e^\epsilon + 1$, it can receive the optimal variance of the aggregation results, which is used to estimate the frequency of each input value reported. Therefore, after employing the OLH protocol, one user's input r_u will be r'_u within the domain of g. During the aggregation process, the aggregator or database can estimate the frequency of each value r'_u in the range F occurs from the following equation,

$$r_u(f) = \frac{\sum_{u=1}^{U} I_f(r'_u) - Nq^*}{p^* - q^*}, \tag{2.6}$$

where $\sum_{u=1}^{U} I_f(r'_u)$ is the counts of occurrence of each value r'_u in the range F, p^* is equal to p from (2.5) is and $q^* = \frac{1}{g}$ (detailed description is in [9]).

2.2 Big Data Analysis : Data-driven Methodology Preliminaries

2.2.1 ζ-structure metric

We use a ζ-structure probability metric, which is a distribution distance measurement proposed in [10, 11] to quantify the distance of distributions. Specifically, a predefined distance measure $d(\mathbb{P}_0, \mathbb{P})$ is constructed on confidence set $\mathcal{D}$, where $\mathbb{P}$ is the true distribution and $\mathbb{P}_0$ is the reference distribution conducted from historical data. The distance d_ζ and confidence set $\mathcal{D}$ can be defined as

$$\mathcal{D} = \{\mathbb{P} : d_\zeta(\mathbb{P}_0, \mathbb{P}) \leq \theta\} \text{ and} \tag{2.7}$$

$$d_\zeta(\mathbb{P}_0, \mathbb{P}) = \sup_{h\in\mathcal{H}} \left| \int_\Omega h d\mathbb{P}_0 - \int_\Omega h d\mathbb{P} \right|. \tag{2.8}$$

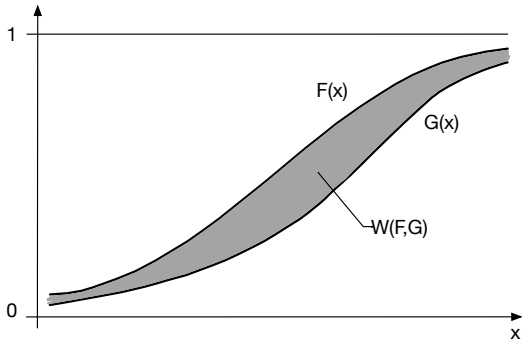

Figure 2.1: Wasserstein metric (one-dimensional case).

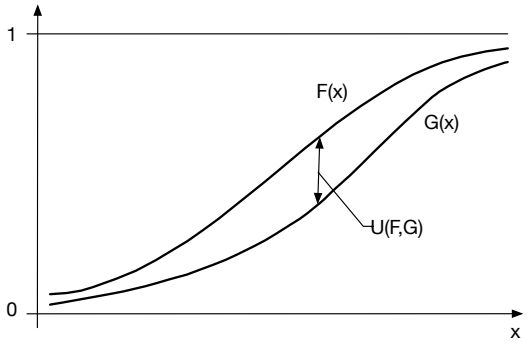

Figure 2.2: uniform metric

Here, $d_\zeta(\cdot,\cdot)$ represents the distance under ζ structure probability metric, θ denotes the tolerance, and $\mathcal{H}$ is a family of real-valued bounded measurable functions on Ω (the sample space on ξ). Tolerance θ is correlated to data size Q, i.e., the size of historical data. It can be easily inferred that the more demand samples that the STED can collect, the tighter $\mathcal{D}$ would be, and the closer ambiguous distribution $\mathbb{P}_0$ would be to $\mathbb{P}$. More details of ζ-structure probability metric is introduced in the next Section.

Three ζ-probability metrics are employed to solve the proposed problem, which are derived as follows. We define $\rho(x,y)$ as the distance between two variables x and y. $\mathbb{P} = \mathcal{L}(x)$ as random variables x following distribution $\mathbb{P}$.

- **Kantorovich metric**: denoted as $d_K(\mathbb{P}_0,\mathbb{P})$, $\mathcal{H} = \{h : ||h||_L \leq 1\}$, where $||h||_L := \sup\{h(x) - h(y)/\rho(x,y) : x \neq y \text{ in } \Omega\}$. By the Kantorovich-Rubinstein theorem, the Kantorovich metric is equivalent to the Wasserstein metric. In particular, when $\Omega = R$,

let d_w denote the Wasserstein metric, then

$$d_w(\mathbb{P}_0, \mathbb{P}) = \int_{-\infty}^{+\infty} |F(x) - G(x)| dx, \tag{2.9}$$

where F and G are the distribution function derived from $\mathbb{P}_0$ and $\mathbb{P}$ respectively, which is demonstrated in Fig. 2.1.

- **Fortet-Mourier metric**: denoted as $d_{FM}(\mathbb{P}_0, \mathbb{P})$, $\mathcal{H} = \{h : ||h||_C \leq 1\}$, where $||h||_C := \sup\{h(x) - h(y)/c(x,y) : x \neq y \text{ in } \Omega\}$ and $c(x,y) = \rho(x,y)\max\{1, \rho(x,a)^{p-1}, \rho(y,a)^{p-1}\}$ for some $p \geq 1$ and $a \in \Omega$. Note that when $p = 1$, Fortet-Mourier metric is the same as Kantorovich metric. The Fortet-mourier metric is usually utilized as a generalization of Kantorovich metric, with the application on mass transportation problems.

- **Uniform metric**: denoted as $d_U(\mathbb{P}_0, \mathbb{P})$, $\mathcal{H} = \{I_{(-\infty,t]}, t \in R^n\}$. According to the definition, we have $d_U(\mathbb{P}_0, \mathbb{P}) = \sup_t |\mathbb{P}_0(x \leq t), \mathbb{P}(x \leq t)|$. It is illustrated in Fig. 2.2, where F and G are the distribution functions derived from $\mathbb{P}$ and $\mathbb{P}_0$, respectively.

From the definition of metrics and relationships between metrics under ζ-structure, we can derive the convergence property and convergence rate accordingly. For the uniform metric, the convergence rate can be derived from the Dvoretzky-Kiefer-Wolfowitz inequality [12–14]:

Proposition 1 The convergence rate of the uniform metric for a single dimension case is (i.e., $n = 1$),

$$\mathbb{P}(d_U(\mathbb{P}_0, \mathbb{P}) \leq \theta) \geq 1 - \exp\left(-\frac{\theta^2 Q}{2}\right). \tag{2.10}$$

In [15], the converge rate of the Kantorovich metric is shown below:

Proposition 2 For a general dimension case (i.e., $n \geq 1$),

$$\mathbb{P}(d_K(\mathbb{P}_0, \mathbb{P}) \leq \theta) \geq 1 - \exp\left(-\frac{\theta^2 Q}{2\varnothing^2}\right). \tag{2.11}$$

Therefore we have $\mathbb{P}(\mathbb{P}_0, \mathbb{P}) \leq \theta) \geq 1 - \exp(-\frac{\theta^2}{2\varnothing^2}Q) = \eta$, and $\theta = \varnothing\sqrt{2log(1/(1-\eta))/Q}$.

The relationship among metrics is represented as $d_{FM}(\mathbb{P}_0, \mathbb{P}) \leq \Lambda \cdot d_K(\mathbb{P}_0, \mathbb{P})$, where $\Lambda = \max\{1, \varnothing^{p-1}\}$ and $\varnothing$ is the diameter of Ω. From the relation between the Fortet-Mourier metric and Kantorovich metric with Proposition 2, we can easily derive the convergence rate of other metrics.

Corollary 1 For a general dimension (i.e., $n \geq 1$), we have

$$\mathbb{P}(d_{FM}(\mathbb{P}_0, \mathbb{P}) \leq \theta) \geq 1 - \exp\Big(-\frac{\theta^2 Q}{2\varnothing^2\Lambda^2}\Big). \tag{2.12}$$

With the convergence rate in (2.10)-(2.12), we can calculate the tolerance θ accordingly. For instance, in the Kantorovich metric, we assume the confidence level is η. Therefore, $\mathbb{P}(d_u(\mathbb{P}_0, \mathbb{P}) \leq \theta)) \geq 1 - \exp(-\frac{\theta^2}{2\varnothing^2}Q) = \eta$ according to (2.10), and $\theta = \varnothing\sqrt{2log(1/(1-\eta)/Q)}$.

Chapter 3

Optimization Based Primary Users' Operational Privacy Preservation

3.1 Introduction

In recent years, the exploding increase of mobile wireless devices and the proliferation of wireless services have accelerated the growth in demand for radio spectrum [16–18]. With limited unlicensed spectrum, regulators are turning to dynamic spectrum sharing and looking for advanced techniques to improve spectrum utilization. As one promising technology, cognitive radio (CR) [19–21] allows secondary users (SUs) to access the idle spectrum in temporal and spatial domain opportunistically, when primary users (PUs) are not active. To further meet the ever-increasing demand for spectrum, Federal Communication Commission (FCC) and National Telecommunications and Information Administration (NTIA) have agreed to open up the 3550-3700 MHz band for unlicensed communications [22, 23]. Note that most frequencies within 3550-3700 MHz are traditionally used by government agencies, e.g., Department of Defense [23,24], and the operational information (such as time of use, geographical locations, anti-jamming capability, and so on) of government facilities, e.g., military radars, are very sensitive or even classified. Therefore, maintaining the PUs' operational privacy while providing SUs' spectrum accessing opportunities poses great challenges.

3.1.1 Related Work

There are several pioneering works about PUs' privacy preservation in existing literature. For example, Clark et al. in [25] discuss several attack models and PUs' obfuscation strategies, based on the assumption that all the information of PUs and SUs are stored in a database. The adversary might hack the database or compromise SUs' devices to infer PUs'

location information. Robertson et al. in [26] proposed to add false spectrum allocation entries into the database to prevent the adversary from learning the operational privacy of PUs. Bahrak et al. in [27] use obfuscation methods to develop a pentagon-shaped contour, which envelops the PU's actual contour to hide PU's accurate location. Another approach is to perturb the output with noises to satisfy differential privacy, as proposed by Dwork et al. [6]. Since simply adding noise signals may degrade the performance of collaborative sensing results, Gao et al. in [28] further proposed a distributed dummy report injection protocol, which jointly prevents the pollution of the aggregation results and preserves location privacy of PUs. Based on attributed-based encryption techniques, Liu et al. in [29] developed the query policy for PUs' spectrum usage database to protect PUs' location privacy. In military communications, Fu et al. in [30] proposed a method that hides traffic characteristics from eavesdroppers by padding the traffic with constant/variable interarrival times, to mitigate the traffic analysis attacks. In addition, there are some previous works related to the time-based traffic model. For instance, Bonal et al. in [31] show that if the underlaying scheduler is fair, the flow-level ('TCP') throughput and delay admit simple time based form, which is independent of the actual inter-arrival distribution between MAC layer packets. However, most existing schemes do not consider the privacy of temporal information such as the time of usage, which are critical for PUs. The temporal operations of PUs might include highly confidential or even classified information (e.g., the operational time of military radars). If such information is obtained by a malicious party, it may jeopardize national security and people's safety. In addition, most of the existing PUs' privacy preserving designs have limited consideration on creating more accessing opportunities to satisfy SUs' traffic demands and improve spectrum utilization, that is the sole purpose of opening up 3550-3700 MHz band for CR communications.

In addition, there are a lot of existing literature works on primary user activity modeling and primary user activity measurement campaigns. For instance, Chen et al. in [32] and Saleem in [33] introduce various spectrum occupancy models which extract different sta-

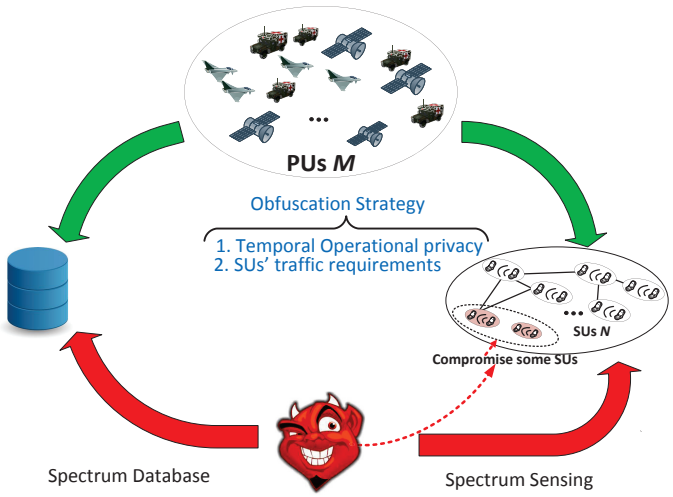

Figure 3.1: System architecture and temporal operational attacks.

tistical properties from the measured data, and discuss the spectrum occupancy prediction which employs moving-average models to predict the channel status at future time instants. Xing et al. in [34] takes the survey of prediction technique in cognitive radio network (i.e., hidden markov model-based prediction, multilayer perceptron neural-network-based prediction, etc.), and present that relevant open research challenges. Hoyhtya et al. in [35] introduce a method to analyze spatial occupancy in location probability metric, and find optimal location for sampling by use of simulated annealing in the article.

From the aspect of the PU, if the PU could precisely predict SU's traffic demands, it can provide better obfuscation strategy. In this way, the PUs can intentionally add dummy signals to obfuscate the attackers[1] while trying their best to satisfy SUs' traffic demands. However, it is a challenging problem to characterize the uncertainty of SUs' traffic demands. Some previous efforts tried to employ robust optimization to address this issue. For instance, Lunden et al. in [36] proposed a robust computationally nonparametric cyclic correlation estimator, which does not require the distribution information of users' traffic. Gong et al. in [37] designed an algorithm to search the optimal detection bound considering signal uncertainty. However, the robust optimization approach can be very conservative, since its

[1]It refers to the attackers either hacking into the spectrum usage database or employing multiple SUs to sense in order to learn the PUs' operational parameters [25].

objective is to minimize the worst case cost or the worst case effectiveness. If PUs add too many dummy signals, according to the overly conservative analysis for privacy preservation, it would reduce the utility of SUs.

3.1.2 Our Contribution

To address these issues, we propose a novel PUs' obfuscation strategy design by formulating the PUs' operational privacy preservation problem as a data-driven risk-averse optimization, and provide robust solutions. Our salient contributions are summarized as follows:

- We introduce a new privacy preserving framework for PUs' obfuscation strategy design, which jointly considers PUs' operational privacy in the temporal domain, the obfuscation cost of PUs, the uncertainty of SUs' demands, and SUs' traffic demand satisfaction under frequency reuse network. Under such a framework, when PUs add dummy signals to obfuscate the adversary, they also need to consider the trade-off between preserving PUs' temporal privacy and satisfying SUs' traffic requirements, and thus cannot arbitrarily generate dummy signals for privacy preserving purposes.

- Under the proposed framework, with abundant historical data of SUs' traffic demands, we allow the PUs to employ data-driven modeling to characterize the uncertainty of SUs' traffic demands. The PUs can build a reference SUs' demand distribution from the historical data, and generate the predicted SUs' demand distribution close to the reference distribution at a certain confidence level. To realize the spectrum reuse under the proposed network, we employ a conflict graph to characterize the transmission interference between SUs, mathematically describe the channel interference relationship between SUs, and employ approximation algorithm to find a sufficiently large number of maximal independent set.

- Based on the modeling of SUs' uncertain traffic demands and temporal operational

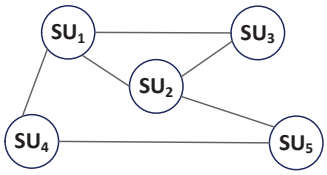

Figure 3.2: A toy overall conflict graph observed by a PU.

privacy metrics, we formulate the PUs' temporal privacy preservation problem into a risk-averse two-stage stochastic optimization under spectrum reuse. We develop algorithms for robust solutions, and conduct simulations to verify our theoretical analysis.

The rest of the work is organized as follows. In Sec. 3.2, we introduce the network model and introduce the related model in the system. In Sec. 3.3, we formulate the PUs' and SUs' utility function, and an optimization problem to preserve PUs' operational privacy. In Sec. 3.4, we develop the solutions to the proposed problem. Simulation results and discussions are presented in Sec. 3.5, and the conclusion remarks are drawn in Sec. 3.6.

3.2 System Description

3.2.1 Network Configuration

As shown in Fig. 3.1, we consider a CR network [38] consisting of N SU transmission pairs, $\mathcal{N} = \{1, 2, \cdots, i, \cdots, N\}$ and M radars (PUs) transmission pairs, $\mathcal{M} = \{1, 2, \cdots, j, \cdots, M\}$, transmitting over non-overlapping brands from 3550-3770 frequency range. Following the principles of overlay CR network communications [39, 40], SUs can opportunistically use the band when the PU owning that band is not active, and SUs must evacuate if the PU comes back. Here, we assume each PU is licensed to use a dedicated band, and each SU can only opportunistically access one band at a time. To preserve temporal operational privacy, PUs will send obfuscating dummy signals periodically, where the fixed period is denoted by $\mathcal{T}$. Let T_j represent the actual temporal spectrum availability for band j (i.e., available time for SUs' opportunistic spectrum accessing before PU j adds dummy

signals), and y_j ($y_j \leq T_j$) be the transformed temporal spectrum availability for band j (i.e., the available time for SUs' opportunistic spectrum accessing after PU j adds dummy signals). Given the transmission rate, let a random variable $d_i(\xi)$ denote the required time to deliver the uncertain traffic demands of SU i within $\mathcal{T}$ corresponding to scenario ξ. For simplicity, we call $d_i(\xi)$ the demand of SU i in the rest of this work, and let $\mathbb{P}_i$ be the distribution of $d_i(\xi)$. For instance, $\mathcal{T} = 60$ mins, and PU j is actively using band j for 20 mins, so that T_j is equal to 40 mins. After PU j executes obfuscation strategy, $y_j = 30$ mins, and the demand of SU i is $d_i(\xi) = 25$ mins.

3.2.2 Other Related Model in the System

3.2.2.1 SU's Transmission Range/Interference Range

When primary services are not active over a certain band, SUs can transmit with full power over that band. Suppose all SUs have the same full transmission power P. The power propagation gain [41] is

$$g_i = \gamma \cdot d_i^{-\alpha} \quad (i \in \mathcal{N}), \tag{3.1}$$

where α is the path loss factor, γ is an antenna related constant, and d_i is the distance between transmitter and receiver of SU pair i[2]. We assume that the data transmission is successful only if the received power at the SU pair's receiver exceeds the receiver sensitivity, i.e., a threshold P_{Tx}. Meanwhile, we assume interference becomes non-negligible only if it is over a threshold of P_{In} at the SU pair's receiver. Thus, the transmission range for a SU is $R_{Tx} = (\gamma P/P_{Tx})^{1/\alpha}$, which comes from $\gamma \cdot (R_{Tx})^{-\alpha} \cdot P = P_{Tx}$. Similarly, based on the interference threshold $P_{In}(P_{In} < P_{Tx})$, the interference range for a SU is $R_{In} = (\gamma P/P_{In})^{1/\alpha}$. It is obvious that $R_{In} > R_{Tx}$ since $P_{In} < P_{Tx}$. Typically, the interference range is 2 or 3 times of the transmission range [41], i.e., $\frac{R_{In}}{R_{Tx}} = 2$ or 3. These two ranges may vary with frequency. The conflict relationship between two SU pairs over the same

[2]The capacity formulation is similar if we consider fading. The major procedure of proposed algorithms will not be changed.

frequency band can be determined by the specified interference range. In addition, if the interference range is properly set, the protocol model can be accurately transformed into the physical model.

3.2.2.2 Conflict Graph

We introduce a conflict graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ to characterize the interference relationship between SUs in the CR network. Following the definitions in [42], we interpret the SU network as a two-dimensional resource space, with dimensions defined by the set of SUs, and the set of available bands. In $\mathcal{G}(\mathcal{V}, \mathcal{E})$, each vertex corresponds to a SU opportunistically accessing a certain band, i.e., a SU-band pair (i, k), where $i \in \mathcal{N}$ and $k \in \mathcal{M}$ [42]. Each SU i stands for a SU transmission pair, including a SU transmitter and a SU receiver from the same SU. Moreover, the distance between transmission pairs is much larger than the distance between transmitter and receiver of SU communication.

Similar to the interference conditions in [41], there is interference if either of the following conditions is true: (i) if two different SUs are using the same band, the receiver of one SU transmission pair is in the interference range of the transmitter in the other SU pair; (ii) a SU pair transmits over two or more bands at the same time. Here, the first condition represents co-band interference, and the second condition represents the radio interface conflicts of SU itself, i.e., the single radio of SU transmitter/receiver cannot support multiple transmissions over multiple bands simultaneously. If there are co-band interferences as shown in the toy conflict graph in Fig. 3.2, we connect two vertices in $\mathcal{V}$ with an undirected edge in $\mathcal{G}(\mathcal{V}, \mathcal{E})$.

Given $\mathcal{G}(\mathcal{V}, \mathcal{E})$, we describe the impact of vertex $i \in \mathcal{V}$ on vertex $j \in \mathcal{V}$ as

$$\delta_{ik} = \begin{cases} 1, \text{ if there is an edge between vertex } i \text{ and } k, \\ 0, \text{ if there is no edge between vertex } i \text{ and } k, \end{cases} \tag{3.2}$$

where two vertices correspond to two SU-band pairs, respectively.

To be more specific, in Fig. 3.2, vertices (SU 1) and (SU 2) stand for SU 1 and SU 2 observed by a PU. They are connected by an edge, which corresponds to the interferences discussed previously. Vertices SU 1 and SU 2 connected through an edge means SU 1 and SU 2 cannot transmit traffic over the spectrum of the PU simultaneously.

3.2.2.3 Maximal Independent Set

Provided that there is a vertex set $\mathcal{I} \subseteq \mathcal{V}$ and a SU-band pair $i \in \mathcal{I}$ satisfying $\sum_{k \in \mathcal{I}, k \neq i} \delta_{ik} < 1$, the transmission at SU-band pair i will be successful even if all the other SU-band pairs in the set $\mathcal{I}$ are transmitting at the same time. If any $i \in \mathcal{I}$ satisfies the condition above, we can reuse the spectrum frequency, and allow the transmission over all these SU-band pairs in $\mathcal{I}$ to be active simultaneously. Such a vertex/SU-band pair set $\mathcal{I}$ is called an independent set. If adding any one more SU-band pair into an independent set $\mathcal{I}$ results in a non-independent one, $\mathcal{I}$ is defined as a maximal independent set (MIS) [42].

3.2.3 Attack Model

In this work, we consider passive adversaries, who may learn the operational time of PUs either from spectrum database or from collective spectrum sensing results of compromised SUs. The compromised SUs do not intercept or modify the messages sent by PUs. Specifically, the adversaries can either eavesdrop the communication between the spectrum database server and SUs, or send queries to the database to learn spectrum availability in the database-driven approach [25,27], or compromise some SUs' devices and collect spectrum sensing[3] results to infer PUs' operational characteristics in the spectrum sensing approach [25], as shown in Fig. 3.1.

[3]Here, we assume SUs use energy detection for spectrum sensing.

3.3 Obfuscation Strategy and Problem Formulation

3.3.1 Utility Functions of PUs and SUs

From the PU's perspective, to preserve the temporal operational privacy from passive attackers, the PU executes obfuscation strategy by generating dummy signals for a certain time period when it actually has no traffic. As a result, the adversary cannot distinguish dummy signals from true signals, by database or collective spectrum sensing. Thereafter, the adversary would obtain transformed temporal spectrum occupation of the PUs based on detected signals, which is a combination of the dummy and true signals. As long as the dummy signals are sent frequently, the PUs' true operations can be hidden in those signals and the operational privacy of PUs can be preserved. Thus, the utility function of PUs' operational privacy preservation can be written as

$$U_{PU_j}(y_j) = c(T_j - y_j), \tag{3.3}$$

where c is a temporal privacy coefficient, T_j is the actual spectrum availability, and y_j is the transformed spectrum availability after the PU j's obfuscation strategy is executed. We can see that if $(T_j - y_j)$ is sufficiently large, PUs' temporal operational privacy is preserved effectively.

From the SUs' perspective, they attempt to transmit on available spectrum to satisfy their own demand. Since SUs can only observe the transformed spectrum availability of PUs, i.e., the spectrum availability after PUs execute obfuscation strategy, we denote the transformed spectrum availability for SU i over spectrum band j as x_i^j. Assuming the SU's traffic can be perfectly split, we let $\sum_{j=1}^{M} x_i^j$ denote the total available time that SU i can transmit over all spectrum bands. We define $d_i(\xi)$ as the actual time needed to satisfy the traffic demand of SU i. Then, $\min\left(\sum_{j=1}^{M} x_i^j, d_i(\xi)\right)$ represents the traffic delivery time of SU

i. Specifically, when $d_i(\xi) < \sum_{j=1}^{M} x_i^j$, which indicates that the time for delivering the traffic demand is less than the transformed available spectrum supply. Then SU i will only in transmit $d_i(\xi)$ to meet its service demands. On the other hand, if transformed available spectrum supply for SU i is less than its real demand, i.e., $\sum_{j=1}^{M} x_i^j < d_i(\xi)$, then SU i will deliver in $\sum_{j=1}^{M} x_i^j$. The utility function of SU i is $U_{SU_i}(d_i(\xi)) = bE_{\mathbb{P}_i}\Big(\min\Big(\sum_{j=1}^{M} x_i^j\omega_i^j, d_i(\xi)\Big)\Big)$. In the network model, the SUs who do not interfere with each other can deliver the traffic on the same spectrum simultaneously. Let ω_i^j denotes the accessing status of SU $i \in \mathcal{N}$ to band $j \in \mathcal{M}$, where $\omega_i^j = 1$ indicates that SU i is opportunistically transmitting over band k, otherwise 0. Given $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ constructed from conflict graph, suppose we can list all MISs as $\mathscr{I} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_q, \cdots, \mathcal{I}_Q\}$, where Q is $|\mathscr{I}|$, and $\mathcal{I}_q \subseteq \mathcal{V}$ for $1 \leq q \leq Q$. Based on the definitions, assumptions and mathematical representations of interference relationship among SUs above, the maximization optimization of utility of function of SUs can be formulated as

$$\max_{x,\omega} \quad b\sum_{i=1}^{N} E_{\mathbb{P}_i}\Big(\min\Big(\sum_{j=1}^{M} x_i^j\omega_i^j, d_i(\xi)\Big)\Big), \tag{3.4}$$

$$\omega_i^j \in \{0,1\}, \qquad (i \in \mathcal{N}, j \in \mathcal{M}), \tag{3.5}$$

$$\sum_{j\in\mathcal{M}} \omega_i^j \leq 1, \qquad (i \in \mathcal{N}), \text{ and} \tag{3.6}$$

$$\omega_i^j \cdot \omega_k^j = 0, \qquad (i,k \in \mathcal{N}, j \in \mathcal{M}, (i,j) \in \mathcal{I}_u, (k,j) \in \mathcal{I}_v, \mathcal{I}_u, \mathcal{I}_v \in \mathscr{I} \text{ and } u \neq v) \tag{3.7}$$

where ω_i^j is optimization variable, b is the SUs' traffic delivery coefficient when SU i is given, and traffic demand $d_i(\xi)$ follows the distribution $\mathbb{P}_i$. Here, binary value ω_i^j indicates the accessing status of SU i to band j, (3.6) means that SU i can only access one band at a time due to the radio interference, and (3.7) presents the SUs who interfere with each other

cannot delivery traffic on same band simultaneously.

The optimization above is a mixed-integer linear programming, which is NP-hard to solve. Some previous work proposed random algorithm for MIS search and adopted in the literature [43], which provides a framework to find more MISs with more computation rounds. However, random search algorithm is quite inefficient for a large size MR-MC network, and could result in redundant search (i.e., getting a MIS already found) with high chance. In [44], Li et al. theoretically develop a polynomial heuristic algorithm to compute set of MISs to better cover the critical MISs in the conflict graph. Moreover, in [44], Li et al. solve the multi-dimensional conflict graph in the network to maximize capacity, which is the same as our scenario (The PU needs to find MISs to make decision to accept/reject proposed SUs considering SUs' mutual interference and spectrum reuse). We employ the greedy algorithm in [44] to find out a large number of MISs (e.g., the number is $Z = 10000$) for approximation instead of finding out all the MIS of $\mathcal{G}(\mathcal{V}, \mathcal{E})$, whose complexity is $\mathcal{O}(M^4N^8)$. By employing Z MISs found in $\mathcal{G}$, we can solve the relaxed optimization in (3.4) by commercial solvers such as CPLEX.

3.3.2 PUs' Operational Privacy Preserving Optimization

Based on the utility functions of PUs and SUs, we expect an obfuscation strategy jointly considering PUs' operational privacy preservation and the satisfaction of SUs' uncertain traffic demands. Regardless of the PU's power consumption, generating more dummy signals obviously better protects the PU's operational privacy but reduces the available opportunistic accessing time of the SUs, diminishing the SUs' traffic delivery. Considering the trade-off between PUs' privacy and SUs' utility, we formulate the PUs' obfuscation strategy design into an optimization, a classic two-stage stochastic programming (SP) problem, described as[4]

[4]If consider operational privacy of primary user over different time period, it can be easily extend. Particularly, $y_t + \lambda_t = y_{t+1} + \lambda_{t+1}$.

$$\max_{y,x,\omega} \quad \sum_{j=1}^{M} c(T_j - y_j)$$

$$+ b \sum_{i=1}^{N} \mathbb{E}_{\mathbb{P}_i} \Big(\min \Big(\sum_{j=1}^{M} x_i^j \omega_i^j, d_i(\xi) \Big) \Big), \tag{3.8}$$

$$\text{s.t.} \quad (3.5), (3.6), (3.7),$$

$$T_j - y_j \geq \lambda, \text{ and} \tag{3.9}$$

$$\sum_{i=1}^{N} x_i^j \omega_i^j \leq y_j. \tag{3.10}$$

The function $\min(\cdot, \cdot)$ in (3.8) considers the influence of PUs' obfuscation strategy $\left(\sum_{j=1}^{M} x_i^j \omega_i^j \right)$ on the SUs' traffic delivery time utility. It is not accurate to just let $d_i(\xi)$ denote the SUs' traffic delivery time utility, since the total available time on PU's spectrum may be less than $d_i(\xi)$. The function $\min(\cdot, \cdot)$ in (3.8) returns the smaller value of $\sum_{j=1}^{M} x_i^j \omega_i^j$ and $d_i(\xi)$. The constraint (3.10) indicates that total transmission time for SUs over PU j's spectrum should be less than the total transformed available spectrum supply of PU j. Besides, to preserve PU j's operational privacy, the time period of the sent dummy signals, i.e., $T_j - y_j$, is then required to be larger than a certain predefined privacy threshold λ, which is a constant, as shown in (3.9).

Due to the ambiguity in demand, it is practically difficult to know the actual probability distribution of SUs' demands. In this work, we employ a data-driven approach, i.e., the risk-averse stochastic optimization approach (RA-SP) allowing distribution ambiguity [15], to characterize the uncertainty of SUs' demands. Instead of deriving a true distribution for the unknown parameter ξ, this optimization approach constructs a confident set D, which allows the distribution ambiguity to be within D under a certain confidence level (e.g., 99%). With RA-SP, considering the worst-case distribution, we can reformulate the problem as

$$
\begin{aligned}
\max_{y,x,\omega} \quad & \sum_{j=1}^{M} c(T_j - y_j) \\
& + \min_{\mathbb{P}_i \in D} \sum_{i=1}^{N} b\, \mathbb{E}_{\mathbb{P}_i} \min\Big(\sum_{j=1}^{M} x_i^j \omega_i^j, d_i(\xi)\Big), \\
\text{s.t.} \quad & (3.5), (3.6), (3.7), \\
& T_j - y_j \geq \lambda \text{ and} \\
& \sum_{i=1}^{N} x_i^j \omega_i^j \leq y_j.
\end{aligned}
\tag{3.11}
$$

We use a distance measurement proposed in [10, 11] to quantify the distance between two distributions. Specifically, a predefined distance measure $d(\mathbb{P}_i^0, \mathbb{P}_i)$ is constructed on confident set D, where $\mathbb{P}_i^0$ is the reference distribution estimated from historical data, and $\mathbb{P}_i$ is the ambiguous distribution of SU i. The distance d and confident set D can be defined from (2.7),(2.8):

$$
D = \{\mathbb{P}_i : d_\zeta(\mathbb{P}_i^0, \mathbb{P}_i) \leq \theta\} \text{ and} \tag{3.12}
$$

$$
d_\zeta(\mathbb{P}_i^0, \mathbb{P}_i) = \sup_{h \in \mathcal{H}} \left| \int_\Omega h d\mathbb{P}_i^0 - \int_\Omega h d\mathbb{P}_i \right|, \tag{3.13}
$$

where the distance under ζ-structure probability metric is denoted by $d_\zeta(\cdot, \cdot)$, the tolerance is denoted by θ, and $\mathcal{H}$ is a family of real-valued bounded measurable functions on Ω (the sample space on ξ). Tolerance θ is correlated to historical data size. It can be easily inferred that the more historical data that the PU can observe, the tighter D would be, and the closer the ambiguous distribution $\mathbb{P}_i$ would be to $\mathbb{P}_i^0$. More details of ζ-structure probability metric is illustrated in Sec.2.2.

Table 3.1: The list of notations

Symbol	Definition
$\mathcal{N}$	Sets of SUs
$\mathcal{M}$	Sets of PUs
T_j	Actual temporal spectrum availability for band j
y_i	Transformed temporal spectrum availability for band j
$d_i(\xi)$	Required time to deliver the uncertain traffic demand of SU i
$\mathcal{G}$	Conflict Graph to characterize the interference relationship among SUs
$\mathcal{V}$	Vertex set in conflict graph $\mathcal{G}$
$\mathcal{E}$	Edge set in conflict graph $\mathcal{G}$
$\mathcal{I}$	Independent set in conflict graph $\mathcal{G}$
ω_i^j	binary variable which indicates if SU i deliver traffic on spectrum j
U	Utility of PUs' operational privacy preservation
b	Traffic delivery coefficient
c	Coefficient of temporal privacy coefficient
x_i^j	Transformed spectrum availability for SU i over spectrum band j
$d_i(\xi)$	Actual traffic time demand for SU i corresponding to scenario ξ
$\mathbb{P}_i$	Real distribution of SU i traffic time demand
$\mathbb{P}_i^0$	Reference distribution of SU i traffic time demand
d_ζ	Distance of two distribution under metric ζ
$\mathcal{D}$	Confidence set
η	Confidence level
θ	Tolerance of the distance between two distributions
Ω	The sample space of ξ
$\varnothing$	The dimension of Ω
x_i^j	transformed spectrum availability for SU i over spectrum band j
ω_i^j	binary variable which indicates if SU_i transmit on PU_j

3.4 Risk-Averse Stochastic Programming for Preserving Temporal Operational Privacy

This section is organized as follows. First, we illustrate the construction of the reference distribution $\mathbb{P}_i^0$ for SU i. Then we represent how to determine tolerance θ on the amount of historical data under ζ-structure. Finally we develop algorithms to solve the problem with respect to different probability distance metrics.

3.4.1 Reference Distribution

First, the reference distribution $\mathbb{P}_i^0$ is defined as

$$\mathbb{P}_i^0(x \leq X) = \frac{1}{Q} \sum_{q=1}^{Q} \delta_{d_q^0(\xi)}(x). \tag{3.14}$$

Suppose we use a set of historical data $\{d_1^0(\xi),\ d_2^0(\xi),\ d_3^0(\xi),\ \cdots,\ d_Q^0(\xi)\}$ to estimate the reference distribution $\mathbb{P}_0$. We utilize the empirical distribution of the historical data samples to construct $\mathbb{P}_0$. To be specific, the distribution in (3.14), the indicator variable $\delta_{d_k^0(\xi)}(x)$ is equal to 1 when $d_k^0(\xi) \leq x$, and 0 otherwise. Then the reference distribution data can be represented by its mass probability p_k^0 which is the ratio of the number of historical data samples matching $d_i(\xi)$ and K, since the supporting space is discrete.

3.4.2 Converge Rate under ζ-Probability Metrics

After that, we explore how to solve the problem in (3.11). The sample space is $\Omega = \{\xi^1, \xi^2, \cdots, \xi^Q\}$. The formulation can be simplified as

$$\begin{aligned} \max_{y,x,\omega} \quad & \sum_{j=1}^{M} c(T_j - y_j) \\ & + \min_{p_i^k} \sum_{i=1}^{N} \sum_{k=1}^{K} b p_i^k \min\Big(\sum_{j=1}^{M} x_i^j \omega_i^j, d_i(\xi)\Big), \end{aligned} \tag{3.15}$$

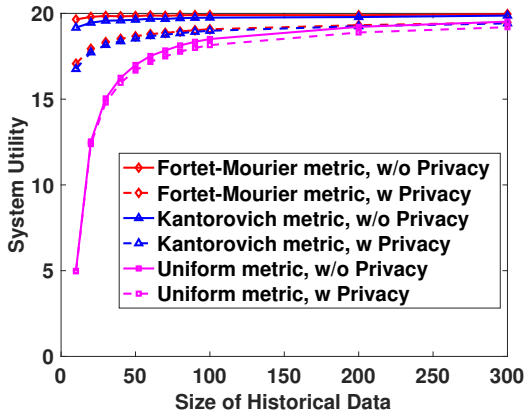

Figure 3.3: Impact of size of historical data on system utility (One SU).

$$\text{s.t.} \quad (3.5), (3.6), (3.7),$$

$$T_j - y_j \geq \lambda, \tag{3.16}$$

$$\sum_{i=1}^{N} x_i^j \omega_i^j \leq y_j, \tag{3.17}$$

$$\sum_{k=1}^{K} p_i^k = 1, \forall i = 1, \cdots, N \text{ and} \tag{3.18}$$

$$\max_{h_k} \sum_{k=1}^{K} h_k p_i^{k0} - \sum_{k=1}^{K} h_k p_i^k \leq \theta, \forall h_k : ||h||_\zeta \leq 1, \tag{3.19}$$

where $|h||_\zeta$ is defined according to different metrics. For the Kantorovich metric and the Bounded-Lipschits metric, $|h_x - h_y| \leq \rho(\zeta^x, \zeta^y)$. For the Fortet-Mourier metric, $|h_x - h_y| \leq \rho(\zeta^x, \zeta^y) \max\{1, \rho(\zeta^x, a)^{p-1}, \rho(\zeta^y, a)^{p-1}\}$. The constraints in (3.18)−(3.19) can be summarized as $\sum_k a_{kl} h_k \leq b_{kl}, l = 1, \cdots, L$. The parameter a_{kl} and b_{kl} is derived from the converge rate relation from different metric shown as (2.10)−(2.12). To reformulate the constraints, we consider the following problem:

$$\min_{h} \quad \sum_{k=1}^{K} h_k p_i^{k0} - \sum_{k=1}^{K} h_k p_i^k \text{ and} \tag{3.20}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} a_{kl} h_k \leq b_{kl}, l = 1, \cdots, L. \tag{3.21}$$

The dual problem can be formulated as

$$\min_{u} \quad \sum_{l=1}^{L} b_l u_l \text{ and} \tag{3.22}$$

$$\text{s.t.} \quad \sum_{l=1}^{L} a_{kl} u_l \geq p_i^{k0} - p_i^k, \forall k = 1, \cdots, V, \tag{3.23}$$

where u is the dual variable. Accordingly, the problem can be reformulated as

$$\max_{y} \quad \sum_{j=1}^{M} c(T_j - y_j)$$

$$+ \min_{p_i^k} \sum_{i=1}^{N} \sum_{k=1}^{K} b p_i^k \min\Big(\sum_{j=1}^{M} x_i^j \omega_i^j, d_i(\xi)\Big), \tag{3.24}$$

$$\text{(SP-M)} \quad \text{s.t.} \quad (3.5), (3.6), (3.7),$$

$$T_j - y_j \geq \lambda, \tag{3.25}$$

$$\sum_{i=1}^{N} x_i^j \omega_i^j \leq y_j \tag{3.26}$$

$$\sum_{k=1}^{K} p_i^k = 1, \sum_{l=1}^{L} b_l u_l \leq \theta, \text{ and} \tag{3.27}$$

$$\sum_{l=1}^{L} a_{il} u_l \geq p_i^{k0} - p_i^k, \forall i = 1, \cdots, N. \tag{3.28}$$

For the Uniform metric, we can have the reformulation from the Uniform metric definition:

$$\max_{y,x,\omega} \quad \sum_{j=1}^{M} c(T_j - y_j)$$

$$+ \min_{p_i^k} \sum_{i=1}^{N} \sum_{k=1}^{K} b p_i^k \min\Big(\sum_{j=1}^{M} x_i^j \omega_i^j, d_i(\xi)\Big) \text{ and} \tag{3.29}$$

$$\text{(SP-U)} \quad \text{s.t.} \quad (3.5), (3.6), (3.7),$$

$$T_j - y_j \geq \lambda, \tag{3.30}$$

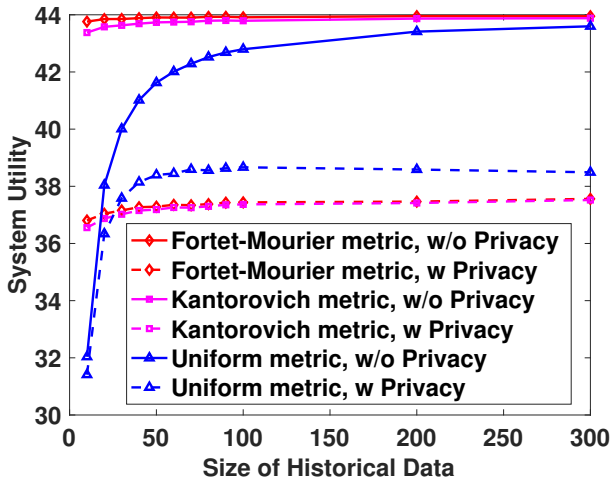

Figure 3.4: Impact of size of historical data on system utility (One SU) under different coefficient c=1, b=4.

$$\sum_{i=1}^{N} x_i^j \omega_i^j \leq y_j, \tag{3.31}$$

$$\sum_{k=1}^{K} p_i^k = 1, \forall i = 1, \cdots, N, \text{ and} \tag{3.32}$$

$$\left| \sum_{k=1}^{l} \left(p_i^{k0} - {p_i}^k \right) \right| \leq \theta, \forall l = 1, \cdots, L. \tag{3.33}$$

The formulation SP-M and SP-U can be solved by CPLEX, etc. We also summarize the algorithm for the problem in Algorithm 3.1, and the detailed description of notation is in Table 3.1.

Algorithm 3.1 Algorithm for Obfuscation Strategy

1: **Input:** Historical data $d_1^0(\xi)$, $d_2^0(\xi)$, $d_3^0(\xi)$ from true distribution. Set η as the confident level of D.
2: **Output:** Objective value of the added time period of dummy signals.
3: Obtain the reference distribution $\mathbb{P}_0^i(x)$ and tolerance θ based on the historical data.
4: Use the reformulation (SP-M) or (SP-U) to solve the problem.
5: Output the solution.

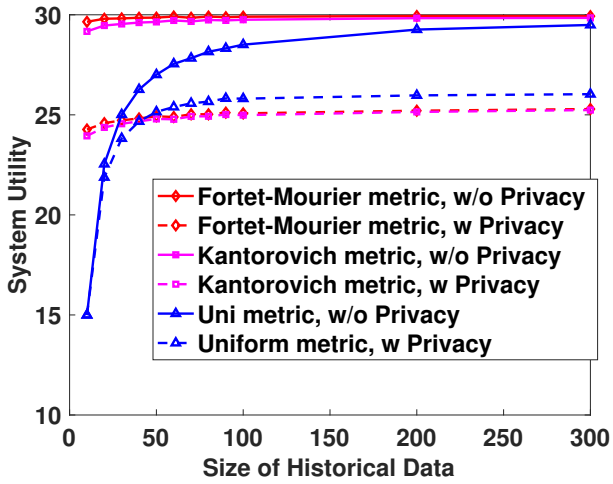

Figure 3.5: Impact of historical data on system utility (One SU) under different distribution.

3.5 Performance Evaluation

For ease of illustration, in the simulations, we consider a CR network of 1 PU and $|\mathcal{N}|$ = 20 SUs, where 20 nodes are randomly deployed in a 1000x1000 m^2 area. Considering the AWGN channel, we assume the noise power σ^2 is 10^{-10} W at all transmitters and receivers. Moreover, we set the path loss factor $\alpha = 4$, the antenna parameter $\gamma = 3.90625$, the receiver sensitivity $P_T = 100\sigma^2 = 10^{-8}$ W and the interference threshold $P_T = 6.25 \times 10^{-10}$ W. We set Z = 10000 as a sufficiently large number for the MISs.

The actual available time of the PU's spectrum is $T = 30$ mins in a particular period $\mathcal{T} = 60$ mins. We set the utility parameter for measuring operational privacy level c to be 3, and the utility parameter for SUs' traffic delivery b to be 5. We assume that traffic demand of all SUs follows a discrete distribution with two scenarios: 10 mins and 20 mins with probabilities 0.4 and 0.6, respectively. We use this distribution to generate the historical data set for simulations.

First, we set the confidence level η to be 98% and the size of historical data varying from 100 to 300, to study the impact of the size of historical data. We also consider two strategies while evaluating performance: with privacy obfuscating strategy ($\lambda = 15min$) and

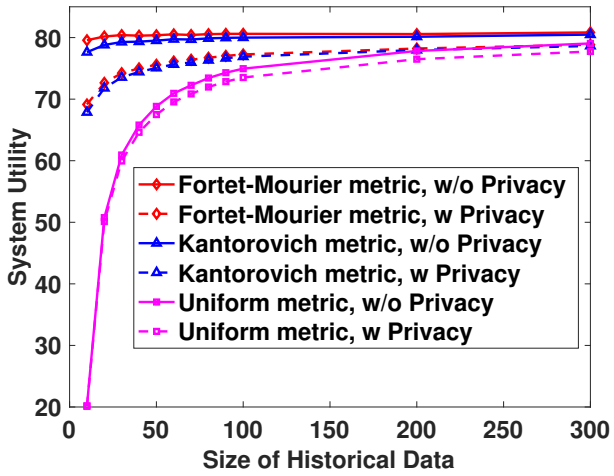

Figure 3.6: Impact of historical data on system utility (10 SUs).

without obfuscating strategy ($\lambda = 0$). First, considering only one SU in cognitive network, the results are reported in Fig. 3.3. From the figure, we can observe that the utility of network increases as the size of historical data increases, irrespective of the kind of metric. The intuition behind the results incurs the value θ decreases as the size of historical data increases. Therefore, the optimized problem in (3.11) becomes less conservative. We can also see that when sample size is 300, the gaps between system utility values are small under all metrics. Moreover, we study the performance under preserving privacy scheme. We set $\lambda = 15$ mins, which indicates that there is at least 15-minute gap between the transformed PUs' spectrum available time and the actual unoccupied period of the PU's spectrum. It can be observed that in Fig. 3.3, the total utility decreases after employing preservation privacy strategy since the PU's operational privacy preservation is at the cost of reducing accessing opportunities for SUs. We can observe that, as the size of historical data increases, the system utility tends to increase under all metric we use. It is because the value of tolerance θ decreases as the number of historical data sample increases, therefore, the risk-averse stochastic problem becomes less conservative. It is shown that the performance under uniform metric is most influenced by the size of historical data, and the performance under Fortet-Mourier metric is always has the highest system utility in the simulation results. In

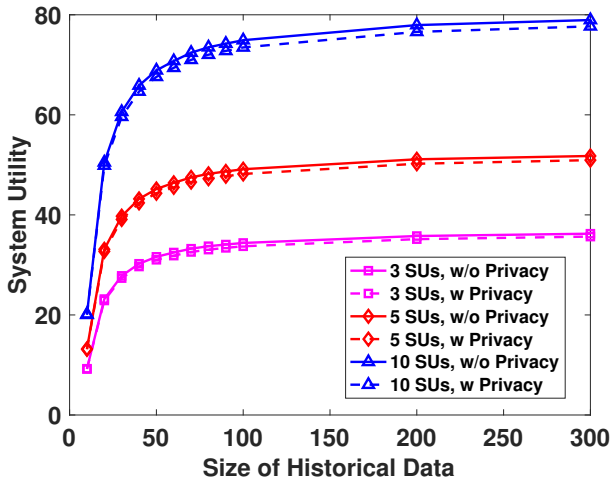

Figure 3.7: Impact of size of historical data and different number of SUs under Uniform metric.

reality, if PUs are very conservative in the predicted distribution of SUs' traffic demand, it should employ uniform metric. On the other hand, the PUs could employ Fourtet-Mourier or Kantorovich Metric to predict the total system utility. To learn the impact on data set and parameter, we set the different value of coefficient ($c = 1$ and $b = 4$), and the different distribution (10 mins and 20 mins with probabilities 0.2 and 0.8). The result is shown in Fig. 3.4 and 3.5. We also have some insights of the system utility under multiple SUs, $|\mathcal{N}|$=10 in Fig. 3.6. We can see that the system utility is much higher after considering frequency reuse in the CR network. To be specific, we compare the system utility under uniform metric for different numbers of SUs in Fig. 3.7. We find that the system utility increases as the number of SUs increases. In Fig. 3.8, we learn the impact of different numbers of SUs under different metrics. It is shown that as the number of SUs increases, the system utility increases, since the size of maximal independent set is larger when more SUs are in the network. Compared to the situation without privacy preserving, the system utility is always lower with privacy preservation scheme under all metrics. Moreover, the system utility under uniform metric has the worst performance.

In addition, we explore the impacts of dummy signals' time period on the system utility. The total number of historical samples is 300, and λ is set from 10 mins to 20 mins,

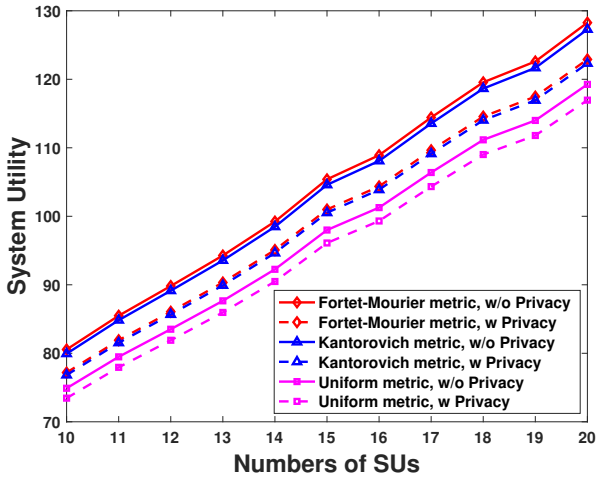

Figure 3.8: Impact of different number of SUs on system utility under different metrics.

and the results are shown in Fig. 3.9. We observe the dummy signal time period increases, the overall system utility under all metrics decreases for chosen PU's privacy coefficient c, SUs' utility coefficient b, and confidence level. The reason is that the contribution of PU's privacy preservation is less important than the deduction of the denied SUs' traffic demands to current system. Also, from Fig. 3.10, we can see that the system utility with more SUs ($|\mathcal{N}| = 5$) in the network is always higher than the system utility with less SUs ($|\mathcal{N}| = 3$) under the same dummy signals time period. However, for a more PU's privacy oriented system (e.g., $c \gg b$), the system utility may increase while adding more dummy signals. For given PUs' and SUs' utility parameters, the proposed scheme can provide a design guideline for such a CR network considering the trade-off between PUs' temporal operational privacy and SUs' performance.

3.6 Conclusion

In this work, we have proposed a novel obfuscation strategy for PUs within 3550-3700 MHz, which has a joint consideration of PUs' temporal operational privacy preservation and SUs' uncertain traffic demands satisfaction under frequency reuse in a cognitive network communication. We have characterized the interference transmission relationship of SUs

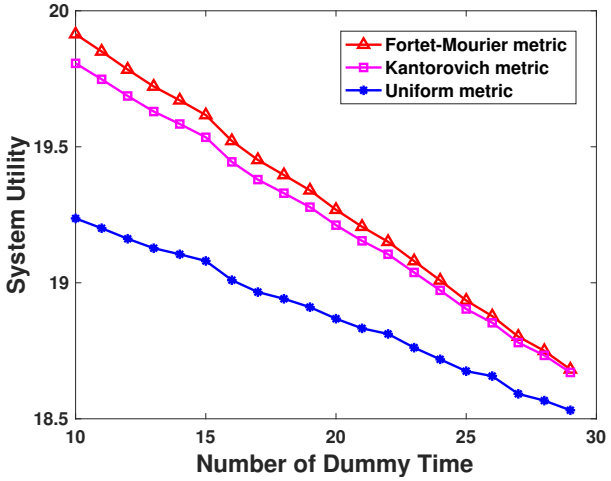

Figure 3.9: PU temporal operational privacy and system utility tradeoff, $|\mathcal{N}| = 1$.

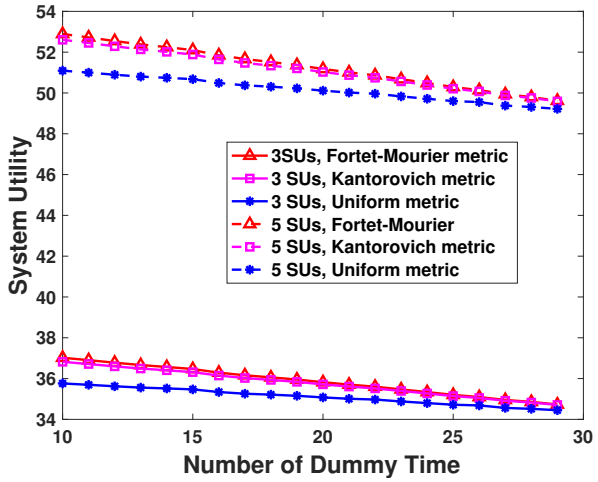

Figure 3.10: Temporal operational privacy and system utility tradeoff, $|\mathcal{N}| = 3$ and $|\mathcal{N}| = 5$.

by constructing conflict graph, and approximation algorithm to find MISs. MoreoverÔºåwe have employed the data-driven risk-averse model in our scheme to characterize SUs' uncertain demand based on the historical data. With such a model, we have formulated the PUs' temporal operational privacy preservation problem into a risk-averse two-stage stochastic optimization. Since the formulated problem is NP-hard to solve, we have relaxed the integer variable and developed a robust algorithm for solutions. Our simulation results to show the effectiveness of the proposed scheme preserving PUs' temporal operational privacy and satisfying SUs' traffic demands. In the future, the research can also be extended as follows.

In our current work we considered the assumption that each SU has only one radio interface, hence each SU transmission pair can only access one PU. In the future, we can study the network model with several radio interfaces for each SU transmission pair. Therefore, each SU can deliver traffic on different spectrums simultaneously. Moreover, by considering a much more complicated distribution of SUs' traffic demand, we are interested in achieving the system utility that better meets the practical circumstances. Finally, we can consider the temporal operational privacy in full duplex communication for CRN according to [45, 46].

Chapter 4

Spectrum Trading with Secondary Users' Differential Privacy Preservation

4.1 Introduction

The last decades have witnessed the proliferation of wireless smart devices, such as smartphones, touchable tablets, intelligent voice assistants (e.g., Amazon Echo or Google Home), etc., and the explosion of various wireless services, which exploit wireless accessing technologies to make people's daily life more convenient and comfortable. Correspondingly, there is a dramatic increase in demand for radio spectrum, while most licensed spectrum bands are underutilized in both temporal and spatial domains [47–49]. Cognitive radio (CR) is a promising technology to improve spectrum utilization, which enables secondary users (SUs) to access the licensed spectrum opportunistically [47–51] when primary users (PUs) are not active. Due to high economic values of spectrum resources, CR technology will potentially initiate spectrum trading, which benefits PUs with monetary gains and SUs with accessing opportunities to satisfy their service demands. Despite those benefits, there are many challenges for pushing spectrum trading in practice. For example, due to hardware limitation of either PUs' or SUs' devices, they may have too limited sensing capability to know some spectrum trading opportunities nearby [50–52]; aiming to maximize the revenue, the PU may feel challenging to develop optimal selling strategies due to the SUs' traffic demand uncertainty; the SU may feel difficult to preserve its spectrum trading privacy (i.e., the SU's locations, true evaluation values of certain spectrum, traffic portfolio, etc.) [3, 27, 53], and so on. Those concerns may make PUs or SUs reluctant to participate in spectrum trading.

To facilitate PUs' and SUs' participation and make spectrum trading practical, recent studies [50–52] have introduced spectrum trading architectures based on existing wireless

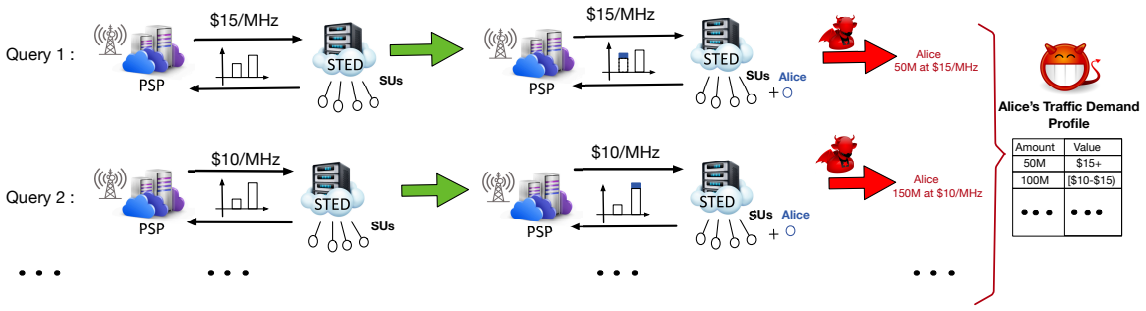

Figure 4.1: Illustrative examples for the traffic demand privacy breach of SUs in spectrum trading.

network infrastructure. Under those architectures, primary service provider (PSP) aggregates vacant spectrum bands from PUs [52], and sells the spectrum bands to secondary service provider (SSP) at wholesale price. The SSP will evaluate the spectrum supply uncertainty [51, 52], make the spectrum purchasing decision, and further sell the purchased spectrum to SUs at retailed price. Here, the role of PSP/SSP can be played by base station in cellular networks, eNodeB in LTE networks, or mobile virtual network operator (MVNO), where the PSP/SSP has more sensing power [51, 52] than the individual SU. Although the spectrum trading architectures in [50–52] help to capture spectrum accessing opportunities, and the algorithms in [50,52] mathematically characterize spectrum supply uncertainty, they ignore the SUs' traffic demand uncertainty, which may have negative impact on PSP's revenue maximization. That is, without the accurate knowledge of SUs' traffic demands, the PSP cannot choose the optimal selling strategies to maximize its revenue. Moreover, the approach of using random variables to model the traffic uncertainty in [50, 52] may be good enough to reflect the PU's traffic patterns over a relatively long-term period, but it will not be able to represent SUs' traffic demands in real-time manner.

Therefore, following the framework of spectrum trading architectures in [50–52], in this work, we further introduce a new entity, called secondary traffic estimator and database (STED), which is responsible for estimating the SUs' traffic demands in real-time manner and answering PSP's queries about SUs' traffic demands as shown in Fig. 4.1. Considering the large population of SUs in the PSP's coverage boundary, it is not efficient to crowdsource

SUs' traffic demands by collecting each SU's demands in terms of time consumption and communication overhead. Thus, we propose to let the STED employ data-driven approach to collect sampled SUs' demands, construct reference demand distribution from sampled demands, and leverage reference distribution to estimate the demand distribution of all SUs.

Now, the leftover challenge hindering spectrum trading is the traffic privacy preservation of the sampled SUs. Taking the query procedure of SUs' demands shown in Fig. 4.1 as an example, the SU's traffic portfolio privacy is breached as follows. For Query 1, the PSP will send a query about SUs' demand to STED, and the query is what the SUs' demand distribution is, if the price for spectrum accessing is \$15/MHz. The STED will respond to this query with a traffic demand distribution of SUs at the cost of \$15/MHz (e.g., 30% SUs would like to purchase 50M and 70% SUs would like to purchase 150M from 100 SUs in total). If a new SU, Alice, joins the group and she would like to purchase 50M at \$15/MHz, the STED will update the SUs' demand distribution to the PSP's query (i.e.,30.7% SUs would like to purchase 50M, 60.3% SUs would like to purchase 150M from 100 SUs in total). From the differences of distributions, the PSP will derive that Alice would like to purchase 50M at \$15/MHz or above. Through multiple queries, the PSP can easily learn Alice's traffic demand profile, which not only discloses Alice's true evaluation values of spectrum resources [54], but also classifies her personal traffic demands (e.g., voice, video, web browsing, social networking, online gaming, etc.) at different price levels.

In order to protect SUs' traffic demand differential privacy (DP) [3,55,56], in this work, we assume the STED is trustworthy, and entitle the STED to transform the SUs' demand distribution by adding noises before it responds to the PSP's queries. Instead of brutally hammering data-driven approach and DP together, we melt SUs' traffic demand DP into data-driven based spectrum trading, and mathematically prove its effectiveness. Based on that, we propose a novel data-driven based spectrum trading scheme with secondary users' differential privacy preservation (3DPP), whose objective is maximizing the PSP's revenue.

Our salient contributions are summarized as follows.

- We propose a novel spectrum trading architecture consisting of the PSP, the SSP, and the STED. Under the proposed architecture, PSP aggregates available spectrum from PUs, and sells the spectrum to the SSP at fixed wholesale price, directly to SUs at spot price, or both as shown in Fig. 4.2. To optimally split the spectrum sold to SSP/SUs, the PSP sends queries to the STED to estimate SUs' demands. The STED will jointly employ data-driven approach and DP preserving techniques to choose sampled SUs, collect their traffic demands, and respond to the PSP's queries.

- We propose a novel 3DPP spectrum trading scheme, which entitles the STED to construct reference distribution $\mathbb{P}_0$ from sampled SUs' demands via data-driven approach. We employ data-driven risk-averse modeling to characterize the uncertainty of SUs' traffic demands, and ensure the uncertainty distance between the reference distribution $\mathbb{P}_0$ and the real traffic demand distribution of all SUs $\mathbb{P}$ is close enough. Besides, we let the STED add noises drawn from Laplace distribution to $\mathbb{P}_0$, and further establish a SUs' traffic demand reference distribution under ϵ-DP, $\mathbb{P}_0'$.

- We mathematically prove that the 3DPP scheme is able to preserve the sampled SUs' traffic demands under ϵ-DP, the references distribution under ϵ-DP, $\mathbb{P}_0'$, and real distribution $\mathbb{P}$ satisfy the data-driven requirements, and the uncertainty distance between the two distributions is close enough, i.e., $\mathbb{P}(d_k(\mathbb{P}_0', \mathbb{P} \leq \theta)) \geq 1 - \exp(-\frac{\theta^2}{2\varnothing^2}V + V\epsilon)$ for Kantorovich metric. Similar proof is applicable for other distribution distance metrics[1].

- Based on the modeling above, we formulate the PSP's revenue maximization into a risk-averse two-stage stochastic problem (RA-SP). To resolve the problem, we utilize ζ-structure probability metric to construct confidence set, and convert the problem into a traditional two-stage robust optimization. We develop algorithms for feasible

[1]Please refer to Sec. 4.4 for details

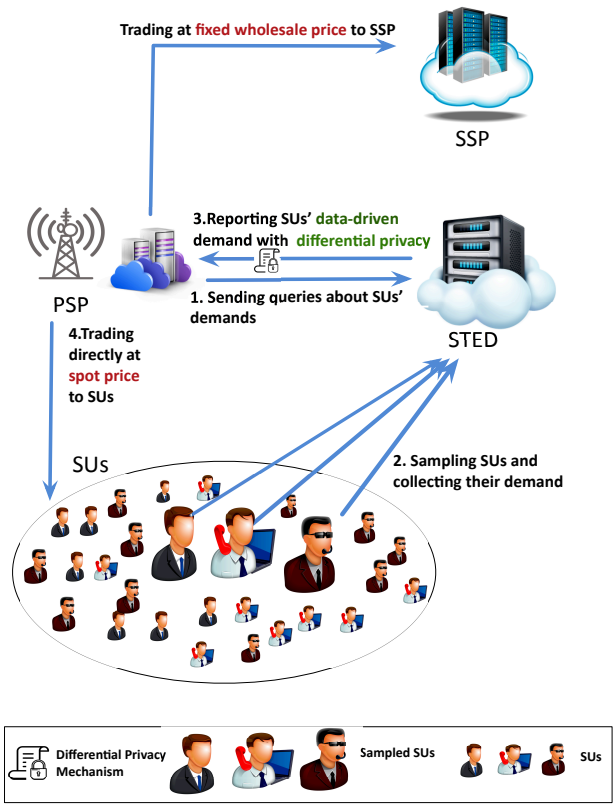

Figure 4.2: The spectrum trading procedure of 3DPP.

solutions and verify the effectiveness of the proposed 3DPP by simulations.

4.2 System Description and 3DPP Outline

4.2.1 System Model and Adversary Model

Our proposed spectrum trading market consists of the PSP, the STED, the SSP, and $\mathcal{N} = \{1, 2, \cdots, i, \cdots, N\}$ SUs as shown in Fig. 4.2. As introduced in Sec. 4.1, the PSP and the SSP are entities similar to MVNOs, and the STED is a trustworthy database server for SUs, which can collect the traffic demand information from SUs, and temporarily store it. The PSP is entitled to aggregate vacant spectrum resources from $\mathcal{M} = \{1, 2, \cdots, j, \cdots, M\}$ PUs with unequal sized bandwidth $\mathcal{W} = \{W_1, W_2, \cdots, W_j, \cdots, W_M\}$, and sell those available spectrum bands for monetary gains.

Similar to power market/cloud resource market in smart grid/cloud computing sys-

tems, the PSP has the following spectrum trading options: (i) selling available spectrum to the SSP at fixed wholesale price, i.e., c; (ii) selling available spectrum bands to the SUs directly at spot price, i.e., b; or (iii) dividing available spectrum resources and selling to both. Thus, before splitting the spectrum and deciding the selling strategy, the PSP will send queries about SUs' demands to the STED as shown in Fig. 4.1. Due to the large number of SUs within the PSP's coverage, the STED will sample some SUs, build up a reference traffic demand distribution of SUs, and respond to the PSP's queries.

The adversaries could be the dishonest PSP or eavesdropping attackers, who are always monitoring the information exchange between the PSP and the STED. As shown in Fig. 4.1, without enforcing any privacy preserving schemes, the adversaries can easily learn the sampled SUs' traffic demand profiles. That may help the adversaries make some illegal monetary gains, or even launch jamming attacks on some valuable services of chosen SUs. It also makes the SUs reluctant to participate in spectrum trading. The meaning of the notations are shown in TABLE 4.1.

4.2.2 3DPP Outline

To preserve the sampled SUs' DP, it takes four steps for the PSP to sell the available spectrum to SUs at spot price b as shown in Fig. 4.1. Firstly, the PSP sends queries about SUs' demands to the STED. Secondly, STED samples some SUs, and constructs a reference demand distribution $\mathbb{P}_0$ from sampled SUs' demands. The STED needs to ensure the uncertainty distance between the reference distribution $\mathbb{P}_0$ and the real traffic demand distribution of all SUs $\mathbb{P}$ is close enough. Thirdly, the STED adds noises drawn from Laplace distribution which is introduced in Sec2.1.1 to $\mathbb{P}_0$, and establishes a SUs' traffic demand reference distribution $\mathbb{P}'_0$, which achieves ϵ-DP. Meanwhile, the STED needs to guarantee that $\mathbb{P}'_0$ is close enough to $\mathbb{P}$, so that $\mathbb{P}'_0$ satisfies both data-driven and ϵ-DP requirements. Then, the STED responds to the PSP's queries with $\mathbb{P}'_0$. Finally, based on $\mathbb{P}'_0$, the PSP decides how much spectrum needs to be sold to the SUs directly at b, and how much

Table 4.1: The list of notations

Symbol	Definition
$\mathcal{N}$	Sets of SUs
$\mathcal{M}$	Sets of PUs
$\mathcal{W}$	Sets of unequal sized bandwidth
c	Fixed wholesale price that PSP sells to SSP
b	Spot price that PSP sells to STED
$\mathbb{P}$	Real traffic demand distribution of all SUs
$\mathbb{P}_0$	Reference distribution from sampled SUs.
$\mathbb{P}'_0$	Distribution after STED adds noise
ϵ	DP parameter
Δf	l_1 sensitivity of a function f in DP
γ_j	Binary variable to indicate if W_j is assigned to STED
ξ	Random variable of SUs' traffic demands
$\mathcal{D}$	Confidence set
η	Confidence level
d_ζ	Distribution distance under ζ-structure probability metric
θ	Tolerance of the distance between two distributions
V	Number of sampled SUs
Ω	The sample space of ξ
$\varnothing$	The dimension of Ω

spectrum need to be sold to the SSP at c to maximize its revenue.

Following this spectrum trading procedure, in the next section, we formulate the PSP's revenue maximization problem under data-driven and DP constraints, i.e., 3DPP. In Sec. 4.3, we theoretically prove that $\mathbb{P}_0'$ is close enough to $\mathbb{P}$, which means the proposed 3DPP has data-driven and DP properties. We also develop solutions to 3DPP problem in Sec. 4.4.

4.3 3DPP Problem Formulation

In this section, we formulate the PSP's revenue maximization problem under data-driven and DP constraints.

4.3.1 PSP's Revenue Maximization Formulation

Let γ_j be a binary variable indicating if W_j is directly sold to SUs, where $\gamma_j = 1$ if W_j is directly sold to SUs, and 0, otherwise. Thus, the PSP's revenue gained from selling spectrum to the SSP can be written as $\sum_{j=1}^{M} cW_j(1-\gamma_j)$, where $1-\gamma_j$ represents the spectrum sold to the SSP at fixed price c. Besides, let random variable ξ denote the uncertain demands from all SUs, and ξ follows distribution $\mathbb{P}$. Then, $b\Big(\min\big(\sum_{j=1}^{M} W_j\gamma_j, \xi\big)\Big)$ is the PSP's revenue gained from selling spectrum to SUs directly[2]. Here, due to the uncertainty of SUs' demands, if the spectrum supply from the PSP (i.e., the spectrum bands that the PSP decided to sell to SUs directly) is more than SUs' actual total traffic demand, i.e., $\sum_{j=1}^{M} W_j\gamma_j > \xi$, the revenue for the PSP is $b\xi$. Otherwise, if the spectrum supply from the PSP is less than SUs' actual traffic demand, i.e. $\sum_{j=1}^{M} W_j\gamma_j < \xi$, the revenue for the PSP is $b\sum_{j=1}^{M} W_j\gamma_j$.

Putting those two parts together, the PSP's revenue maximization can be formulated as

$$\max_{\gamma} \quad -\sum_{j=1}^{M} cW_j\gamma_j + \sum_{j=1}^{M} cW_j$$

[2]In this work, we assume the aggregated spectrum resources can be perfectly split to satisfy SUs' traffic demands.

$$+ b\mathbb{E}_{\mathbb{P}}\Big(\min\big(\sum_{j=1}^{M} W_j\gamma_j, \xi\big)\Big), \tag{4.1}$$

$$\text{s.t.:} \qquad \gamma_j \in \{0,1\}, j = 1, \cdots, M, \text{ and} \tag{4.2}$$

$$\xi = \sum_{i=1}^{N} d_i, i = 1, \cdots, N, \tag{4.3}$$

where γ_j is binary variable, and (4.3) represents the total traffic demand of all SUs.

4.3.2 Data-Driven Based PSP's Revenue Optimization

Given the huge number of SUs within PSP's coverage, the STED cannot collect traffic demand information from every possible SU, i.e., the STED is generally difficult to obtain the true probability distribution of all SUs' demand $\mathbb{P}$. Instead, we allow the STED to collect the traffic demands from a series of sampled SUs, and construct reference demand distribution $\mathbb{P}_0$. For a given set of sampled SU data, it is easy for us to construct a histogram to fit the SUs' traffic demand. For example, we can set N intervals to fit the total traffic demand of sampled SUs in each interval to be $L_1, \cdots, L_n, \cdots, L_N$ with $L = \sum_{n=1}^{N} L_n$. For instance, L_1 is the number of SUs who would like to access spectrum on price \$15/MHZ, L_2 is the number of SUs who would like to access spectrum on price \$20/MHZ, etc.. Based on this, we can construct an reference distribution for the uncertain total traffic demand of all consumers in particular time period of a day as $p_1^0 = L_1/L, \cdots, p_n^0 = L_n/L, \cdots$, and $p_N^0 = L_N/L$. For simplicity, we let $\mathbb{P}_0 = p_1^0, p_2^0, \cdots, p_N^0$ represent the corresponding reference distribution. Since $\mathbb{P}_0$ may not be 100% represents the unique true SUs' demand distribution $\mathbb{P}$, we employ risk-averse stochastic optimization approaches (RA-SP) allowing distribution ambiguity [15] to reformulate the PSP's revenue maximization problem in (4.1). Instead of deriving a true distribution for ξ, this optimization approach derives a confidence set D, and allows the distribution ambiguity to be within set $\mathcal{D}$ with a certain confidence level (e.g.,

99%). The data-driven based RA-SP for the PSP's revenue maximization is formulated as

$$\max_{\gamma} \quad -\sum_{j=1}^{M} cW_j\gamma_j + \sum_{j=1}^{M} cW_j + \min_{\mathbb{P}\in\mathcal{D}} b\mathbb{E}_{\mathbb{P}}\Big(\min\big(\sum_{j=1}^{M} W_j\gamma_j, \xi\big)\Big), \tag{4.4}$$

$$\text{s.t.:} \quad \text{constraints (4.2) and (4.3).}$$

We use a distribution distance measurement proposed in [10, 11] to quantify the distance of distributions. Specifically, a predefined distance measure $d(\mathbb{P}_0, \mathbb{P})$ is constructed on confidence set $\mathcal{D}$, where $\mathbb{P}$ is the true distribution and $\mathbb{P}_0$ is the ambiguous distribution conducted from sampled SUs. The distance d_ζ and confidence set $\mathcal{D}$ can be defined as (2.7)−(2.8) which is shown as

$$\mathcal{D} = \{\mathbb{P} : d_\zeta(\mathbb{P}_0, \mathbb{P}) \leq \theta\} \text{ and} \tag{4.5}$$

$$d_\zeta(\mathbb{P}_0, \mathbb{P}) = \sup_{h\in\mathcal{H}} \left| \int_\Omega h d\mathbb{P}_0 - \int_\Omega h d\mathbb{P} \right|. \tag{4.6}$$

Here, $d_\zeta(\cdot, \cdot)$ represents the distance under ζ structure probability metric, θ denotes the tolerance, and $\mathcal{H}$ is a family of real-valued bounded measurable functions on Ω (the sample space on ξ). Tolerance θ is correlated to data size, i.e., the number of SUs' demand samples. It can be easily inferred that the more demand samples that the STED can collect, the tighter $\mathcal{D}$ would be, and the closer ambiguous distribution $\mathbb{P}_0$ would be to $\mathbb{P}$. More details of ζ-structure probability metric is introduced in the next Section.

4.3.3 3DPP: Data-Driven Based PSP's Revenue Optimization under ϵ-DP

To protect the sampled SUs' traffic demand profiles, the STED will employ Laplace mechanism to add noises into $\mathbb{P}_0$. Here, we denote $\mathbb{P}_0'$ as the distribution after employing Laplace mechanism, and p_0' as its density of probability function accordingly. According to the definition of ϵ-DP, we have $p_0' \leq p_0 e^\epsilon$. Thus, the data-driven based PSP's revenue

maximization under ϵ-DP, i.e., 3DPP problem, can be reformulated as

$$\max_{\gamma} \quad -\sum_{j=1}^{M} cW_j\gamma_j + \sum_{j=1}^{M} cW_j + \min_{\mathbb{P}\in\mathcal{D}'} b\mathbb{E}_{\mathbb{P}}\Big(\min\big(\sum_{j=1}^{M} W_j\gamma_j, \xi\big)\Big), \tag{4.7}$$

$$\text{s.t.:} \quad (4.2), (4.3)$$

$$\mathcal{D}' = \{\mathbb{P} : d_\zeta(\mathbb{P}'_0, \mathbb{P}) \leq \theta\}, \text{ and} \tag{4.8}$$

$$d_\zeta(\mathbb{P}'_0, \mathbb{P}) = \sup_{h\in\mathcal{H}} \left| \int_\Omega h d\mathbb{P}'_0 - \int_\Omega h d\mathbb{P} \right|. \tag{4.9}$$

4.4 3DPP Proof and Solutions

This section is organized as follows. First, we present how to determine converge rate under ζ-structure probability structure. We show the relation between DP parameter ϵ and distribution tolerance θ in ζ-structure probability structure, and prove our DP mechanism satisfies the requirement of data-driven, which is $d_\zeta(\mathbb{P}'_0, \mathbb{P}) \leq \theta$. Second, we reformulate the problem under ζ-structure probability metrics, and convert it to a traditional two-stage robust optimization. We develop algorithms to solve the problem w.r.t. different probability metrics.

As described in Sec.2.2, we employ three different ζ-structure probability metrics and solve our problem under these constraints correspondingly in (2.10)$-$(2.12).In this work, we further prove the converge rate between distribution with Laplace mechanism $\mathbb{P}'_0$ and real distribution $\mathbb{P}$ under Kantorovich metric as follows.

We define $\rho(x, y)$ as the distance between two variables x and y, and $\varnothing$ as the dimension of Ω. $\mathbb{P} = \mathcal{L}(x)$ represents random variables x follows distribution $\mathbb{P}$. We denote V as the size of sampled SUs.

Proposition 1 For a general dimension case (i.e., n≥1),

$$\mathbb{P}(d_K(\mathbb{P}_0', \mathbb{P}) \leq \theta) \geq 1 - \exp\bigg(- \frac{\theta^2 V}{2\varnothing^2} - \epsilon \bigg). \tag{4.10}$$

Proof. Let us define a set

$$\mathcal{B} := \{\mu \in \mathcal{P}(\Omega) : d_k(\mu, \mathbb{P}) \geq \theta\}, \tag{4.11}$$

where $\mathcal{P}(\Omega)$ is the set of all probability measures defined on Ω. Let $\mathcal{C}(\Omega)$ be the set of bounded continuous function $\phi \to R$. Therefore, following the definitions, for each $\phi \in \mathcal{C}(\Omega)$, we have

$$\mathbb{P}\big(d_K(\mathbb{P}_0', \mathbb{P}) \geq \theta\big) = Pr(\mathbb{P}_0' \in \mathcal{B}), \tag{4.12}$$

$$\leq Pr\bigg(\int_\Omega \phi d\mathbb{P}_0' \geq \inf_{\mu \in \mathcal{B}} \int_\Omega \phi d\mu \bigg), \tag{4.13}$$

$$\leq \exp\bigg(-V \inf_{\mu \in \mathcal{B}} \int_\Omega \phi d\mu \bigg) E\left(e^{V \int_\Omega \phi d\mathbb{P}_0 e^\epsilon} \right), \tag{4.14}$$

$$= \exp\bigg(-V \inf_{\mu \in \mathcal{B}} \bigg\{ \int_\Omega \phi d\mu - \frac{1}{V} \log E\left(e^{V \int_\Omega \phi d\mathbb{P}_0 e^\epsilon} \right) \bigg\} \bigg),$$

$$= \exp\bigg(-V \inf_{\mu \in \mathcal{B}} \bigg\{ \int_\Omega \phi d\mu - \frac{1}{V} \log E\left(e^{\sum_{i=1}^V e^\epsilon \phi(\xi^i)} \right) \bigg\} \bigg), \tag{4.15}$$

$$= \exp\bigg(-V \inf_{\mu \in \mathcal{B}} \bigg\{ \int_\Omega \phi d\mu - \log \int_\Omega e^\epsilon e^\phi d\mathbb{P} \bigg\} \bigg), \text{ and} \tag{4.16}$$

$$= \exp\bigg(-V \inf_{\mu \in \mathcal{B}} \bigg\{ \int_\Omega \phi d\mu - \log \int_\Omega e^\phi d\mathbb{P} - \epsilon \bigg\} \bigg), \tag{4.17}$$

where (4.12) follows the definition of $\mathcal{B}$, inequality (4.13) is from the fact that $\mathbb{P}_0 \in \mathcal{B}$, and μ is the one distribution in $\mathcal{B}$ that satisfies the minimum of $\int_\Omega \phi d\mu$, (4.14) follows from the Chebyshev's exponential inequality [14], and (4.15) follows from the definition of $\mathbb{P}_0$.

Now we define $\Delta(\mu) := \sup_{\phi \in \mathcal{C}(\Omega)} \int_\Omega \phi d\mu - \log \int_\Omega e^\phi d\mathbb{P}$. Thus, following the definition of $\mathcal{C}(\Omega)$, there exists a series ϕ_n such that $\lim_{n\to\infty} \int_\Omega \phi d\mu - \log \int_\Omega e^\phi d\mathbb{P} = \Delta(\mu)$. For any small positive number $\theta' > 0$, there exists a constant number n_0 such that $\Delta(\mu) - (\int_\Omega \phi_n d\mu - \log \int_\Omega e_n^\phi d\mathbb{P}) \leq \theta'$ for any $n \geq n_0$. Therefore, according to (4.17), we use substitute ϕ_n for

Algorithm 4.1 Algorithm1: Procedure of Solving $3DPP$

1: **Input:** Historical data ξ_1,ξ_2,$\cdots$,ξ_N from sample SUs. Set ϵ as the privacy parameter. Set η as the confidence level of D.
2: **Out:** Objective value of the η.
3: STED receives the number of sampled SUs under different traffic demand, i.e.,ξ_1, $\cdots$, ξ_N.
4: STED adds Laplace noise to the original data set of sample SUs. ξ'_n=ξ_n+$(Y_1, \cdots, Y_k)$, where Y_i are i.i.d random variables drawn from Lap$(\Delta f/\epsilon)$.
5: STED reports the processed data ξ'_n to PSP.
6: Obtain the reference distribution $\mathbb{P}'_0(\xi)$ and tolerance θ based on the data received from STED.
7: STED uses the reformulation (SP-M) or (SP-U) to solve the problem.
8: Output the solution.

ϕ, then we have

$$\begin{aligned}
&\Pr(\mathbb{P}'_0 \in \mathcal{B}) \\
&\leq \exp\left(-V \inf_{\mu\in\mathcal{B}} \left\{\int_\Omega \phi d\mu - \log \int_\Omega e^{\phi} d\mathbb{P} - \epsilon\right\}\right) \text{ and} \qquad (4.18)\\
&\leq \exp\left(-V \inf_{\mu\in\mathcal{B}} \left\{\Delta(\mu) - \epsilon - \theta'\right\}\right). \qquad (4.19)
\end{aligned}$$

According to Lemma 6.2.13 in [12], we have

$$\Delta(\mu) = d_{KL}(\mu, \mathbb{P}) \qquad (4.20)$$

where $d_{KL}(\mu, \mathbb{P})$ is the discrete case KL-divergence defined as $\sum_i ln(p_i/\mu_i)p_i$. For the case $\mu \in \mathcal{B}$, with (4.11), we have $d_K(\mu, \mathbb{P}) \geq \theta$. Moreover, in "Particular case 5" in [13], we have

$$d_K(\mu, \mathbb{P}) \leq \varnothing\sqrt{2d_{KL}(\mu, \mathbb{P})} \qquad (4.21)$$

hold for $\forall\mu \in \mathcal{P}(\Omega)$. Consequently, following (4.21), we have

$$d_{KL}(\mu, \mathbb{P}) \geq \theta^2/\left(2\varnothing^2\right). \qquad (4.22)$$

Combining (4.19), (4.20), (4.22), we have

$$\Pr(\mathbb{P}'_0 \in \mathcal{B}) \leq \exp\left(-V\left(\frac{\theta^2}{2\varnothing^2} - \epsilon - \theta'\right)\right). \qquad (4.23)$$

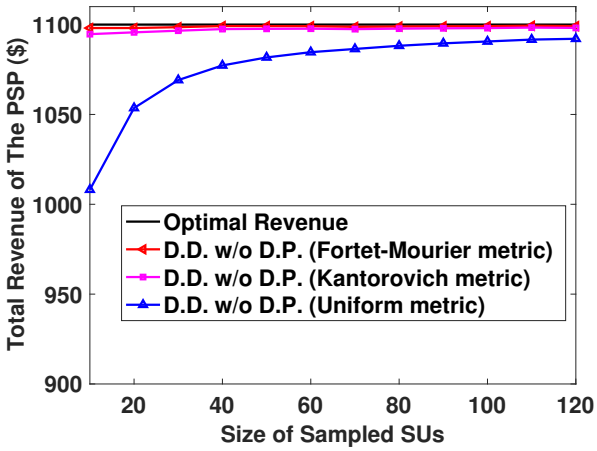

Figure 4.3: Data-Driven spectrum trading without ϵ-DP.

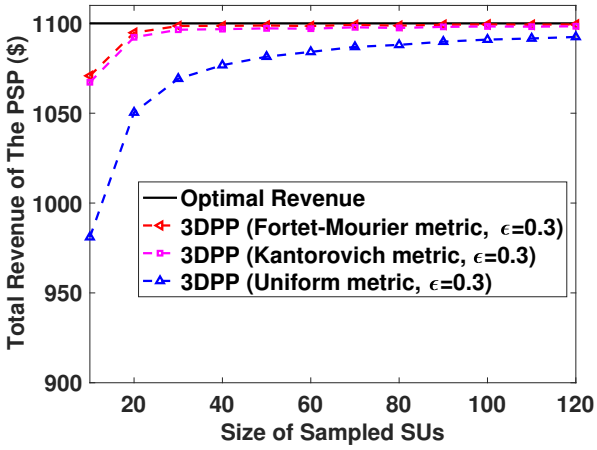

Figure 4.4: Total revenue of PSP under different probability distance metrics.

Let $\theta' = \lambda/V$ for any arbitrary small positive λ. Then, we have

$$\begin{aligned} &\Pr\left(d_k\left(\mathbb{P}_0', \mathbb{P}\right) \geq \theta\right) \\ &= \Pr\left(\mathbb{P}_0' \in \mathcal{B}\right) \leq \exp\left(-V\left(\frac{\theta^2}{2\varnothing^2} - \epsilon\right) + \lambda\right). \end{aligned} \tag{4.24}$$

Since λ can be arbitrarily small, we have $\mathbb{P}(d_k(\mathbb{P}_0', \mathbb{P} \leq \theta)) \geq 1 - \exp(-\frac{\theta^2}{2\varnothing^2}V + V\epsilon)$.

With convergence rate (4.24), we can calculate the tolerance θ accordingly. For instance, in Kantorovich metric, we assume the confidence level is η. Therefore $\mathbb{P}(d_u(\mathbb{P}_0, \mathbb{P} \leq \theta)) \geq 1 - \exp(-\frac{\theta^2}{2\varnothing^2}V + V\epsilon) = \eta$ according to (4.24), and $\theta = \varnothing\sqrt{2log(e^{\epsilon V}/(1-\eta))/V}$. $\square$

Similar proof is applicable for other metrics. For example, following the proof procedure of **Proposition 1** in our work and using Corollary 1 in [15], it is easy to prove that under Fortet-Mourier metric, we have

$$\mathbb{P}(d_{FM}(\mathbb{P}_0', \mathbb{P}) \leq \theta) \geq 1 - \exp\Big(- \frac{\theta^2 V}{2\varnothing^2 \Lambda^2} + \epsilon V \Big), \tag{4.25}$$

where $\Lambda = \max\{1, \varnothing^{p-1}\}$. Due to the page limits, we omit the detailed proof procedure.

4.4.1 Problem Reformulation under ζ-Probability Metrics, and Solutions

We denote $x = \sum_{j=1}^{M} W_j \gamma_j$, $\alpha = \sum_{j=1}^{M} W_j$ where α is a constant. The sample space is $\Omega = \{\xi_1, \xi_2, \cdots, \xi_N\}$. Then the formulation can be simplified as

$$\max_{x} \qquad - cx + \min_{p_i} b \sum_{i=1}^{N} p_i \Big(\min\big(x, \xi_i\big) \Big) + c\alpha \tag{4.26}$$

$$\text{s.t.} \qquad x \in [0, \alpha], \tag{4.27}$$

$$\sum_{i} p_i = 1, \text{ and} \tag{4.28}$$

$$\max \sum_{i=1}^{N} h_i p_{0_i}' - \sum_{i=1}^{N} h_i p_i \leq \theta, \forall h_i : ||h||_\zeta \leq 1, \tag{4.29}$$

where the $|h||_\zeta$ is defined according to different metric. In Kantorovich metric, $|h_x - h_y| \leq \rho(\zeta^x, \zeta^y)$. The constraint (4.28), (4.29) can be summarized as $\sum_i a_{il} h_i \leq b_{il}, l = 1, \cdots, L$. To reformulate the constraint, we consider the problem

$$\min_{h_i} \qquad \sum_{i=1}^{N} h_i p_{0_i}' - \sum_{i=1}^{N} h_i p_i \text{ and} \tag{4.30}$$

$$\text{s.t.} \qquad \sum_{i=1}^{N} a_{il} h_i \leq b_{il}, l = 1, \cdots, L. \tag{4.31}$$

Its dual problem is represented as

$$\min \qquad \sum_{l=1}^{L} b_l u_l, \text{ and} \tag{4.32}$$

$$\text{s.t.} \qquad \sum_{l=1}^{l} a_{il} u_l \geq p'_{0_i} - p_i, \forall i = 1, \cdots, N, \tag{4.33}$$

where u is the dual variable. Accordingly, the formulation can be reformulated as

$$\max_{x} \qquad -cx + \min_{p_i} b \sum_{i=1}^{N} p_i \Big(\min\big(x, \xi_i\big) \Big) + c\alpha, \tag{4.34}$$

$$\text{(SP-M)} \qquad \text{s.t.} \qquad x \in [0, \alpha], \tag{4.35}$$

$$\sum_{i=1}^{N} p_i = 1, \sum_{l=1}^{l} b_l u_l \leq \theta, \text{ and} \tag{4.36}$$

$$\sum_{l=1}^{L} a_{il} u_l \geq p'_{0_i} - p_i, \forall i = 1, \cdots, N. \tag{4.37}$$

For the uniform metric, we can have the reformulation from the Uniform metric definition

$$\max_{x} \qquad -cx + \min_{p_i} b \sum_{i=1}^{N} p_i \Big(\min\big(x, \xi_i\big) \Big) + c\alpha \tag{4.38}$$

$$\text{(SP-U)} \qquad \text{s.t.} \qquad x \in [0, \alpha], \tag{4.39}$$

$$\sum_{i=1}^{N} p_i = 1, \text{ and} \tag{4.40}$$

$$\left| \sum_{i=1}^{l} \left(p'_{0_i} - p_i \right) \right| \leq \theta, \forall l = 1, \cdots, L. \tag{4.41}$$

The formulation SP-M and SP-U can be solved by L-shape algorithm which is described in [57]. We summarize the procedure of solving the 3DPP problem in Alg. 4.1.

4.5 Performance Evaluation

4.5.1 Simulation Setup

For illustrative purposes, we consider a spectrum trading market with 500 SUs. We assume the true traffic demand of all SUs follows a discrete distribution: 100M with probability 0.4 and 200M with probability 0.6, respectively. Total available spectrum resources

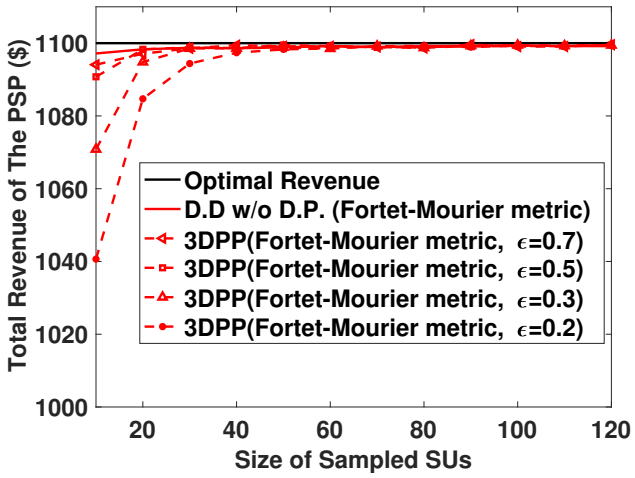

Figure 4.5: Total revenue of the PSP with 3DPP under Fortet-Mourier metric

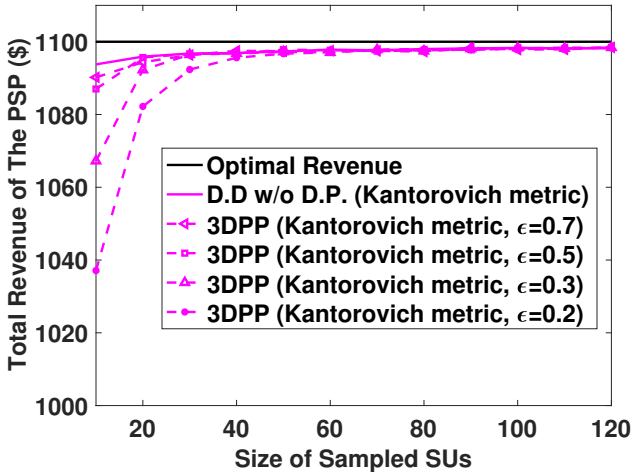

Figure 4.6: Total revenue of the PSP with 3DPP under Kantorovich metric metric

aggregated by the PSP is 300M. In addition, we set the fixed wholesale price for the spectrum sold to the SSP to be $ 3/MHz, and the spot price for the spectrum sold directly to SUs to be $ 5/MHz.

4.5.2 Privacy and Performance Analysis

First, the confidence level η is set to be 90% and the size of sampled SUs varies from 10 to 120. We study the data-driven algorithm without DP. The results are shown in Fig. 4.3. After collecting traffic demand of sample SUs, the STED does not add Laplace noises, and

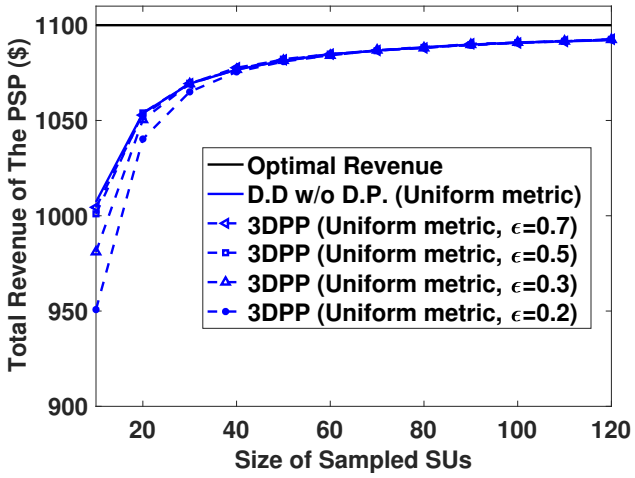

Figure 4.7: Total revenue of the PSP with 3DPP under Uniform metric metric

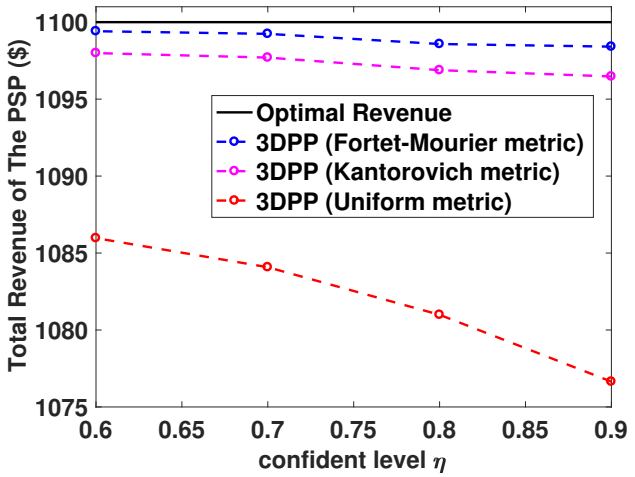

Figure 4.8: Total revenue of the PSP with 3DPP under different confidence levels.

submits the true reference distribution directly to the PSP. From the results in Fig. 4.3, it can be observed the total revenue of the PSP increases when the size of sample SUs increases, regardless of the distance metrics adopted. The intuition behind the result is that, as the size of sampled SUs , the value θ decreases, which stands for the distance between true distribution and reference distribution. As a result, the solutions are moving closer to the optimal one. It is also shown in Fig. 4.3 that the gap between total revenue under the Fortet-Mourier metric and the Kantorovich metric is very small, when the number of sampled SUs is over 100. When the number of sampled SU is 120, the results under all metrics are close to the optimal one. Besides, we study the 3DPP's performance in Fig. 4.4. Compared with results in Fig. 4.3, it can be observed that the total revenue of the PSP with 3DPP is less than that without ϵ-DP when the number of sampled SUs is small, but becomes close to each other, or even to the optimal revenue when the size of sampled SUs increases. That means it incurs some cost to involve ϵ-DP for the sampled SUs' traffic demands, especially when the number of sampled SUs is small. But this impact significantly diminishes when the number of samples increases. That also implies that the proposed 3DPP scheme can still successfully captures the characteristics of whole data set, i.e., the demand distribution of all SUs, while preserving individual sampled SU's traffic profile privacy. Moreover, from Fig. 5, we found the Fortet-Mourier metric is a more applicable metric, since the simulation results is more closer to the optimal revenue.

Moreover, we explore the impact of DP parameter ϵ in Fig. 4.5 - Fig. 4.7. We choose four different ϵ values, i.e., 0.7, 0.5, 0.3, 0.2, respectively, and study its impact under different metrics. We find that as the ϵ decreases, the total revenue of PSP decreases under all metrics. The reason is, ϵ stands for the upper bound of privacy loss. It means, when ϵ is smaller, the mechanism yields better privacy, and less accurate responses which leads to less revenue of the PSP. It also can be observed that, when size of sampled SUs is less than 60, the gaps of total revenue under different ϵ is large. When size of sampled SUs increases, the influence of ϵ is less, and the total revenue under 3DPP with different ϵ converges to the optimal

one. Last but not the least, we study the effect of confidence level on the 3DPP in Fig. 4.8. We set the number of sampled SUs as 40, and test four different confidence levels, i.e., 0.6, 0.7, 0.8, 0.9, respectively. From the Fig. 4.8 we can observe that, as the confidence level increases, the gaps between the PSP's revenue of 3DPP and optimal one increases under all three metrics. The reason is that, as the confidence level η increases, the distance θ between reference distribution with ϵ-DP $\mathbb{P}_0'$ and true distribution $\mathbb{P}$ increases, and the true probability distribution of SUs traffic demands is more likely to be in the confidence set $\mathcal{D}$. That implies that distribution in set $\mathcal{D}$ which is not that close to $\mathbb{P}$ might be used to yield solutions. Therefore, the PSP's revenue performance degrades when confidence level increases.

4.6 Related Work

There are a lot of research works focusing on preserving privacy during spectrum trading. To be specific, Errapotu et al. in [58] employ the Paillier,Äô crypto-system to preserve SUs' bidding privacy and maximize the revenue of PU simultaneously in a semi-distributed manner. Liu et al. leverage attribute-based encryption to preserve PUs' operational privacy in spectrum database. Recently, a promising mechanism, *differential privacy* (DP), proposed by Dwork [55] has been employed in dynamic spectrum allocation [3, 56, 59]. DP aims to reveal statistical information of whole dataset without compromising the privacy of each individual. Zhu et al. in [56] preserve the bidders,Äô valuation privacy with approximate revenue maximization in spectrum auction mechanism, and theoretically proved the mechanism is differential private. In the area of internet of things and spectrum monitoring, Sun et al. in [59] propose a distributed stream monitoring system with high communication efficiency and privacy guarantee. The technique they proposed is powered by DP theory, which can ensure submitted data of every node are not substantially different with one element of the node's data stream changes. Jin et al. in [3] present a crowdsourced spectrum sensing service provider, which selects spectrum-sensing participants in a DP preserving manner.

They prove the new mechanism can prevent any internal or external attackers from learning the location of mobile participants, and minimize the social cost simultaneously.

To process spectrum trading, PU service provider recruits SUs to collect their characteristic (traffic demand, location, etc.), and allocate different quantity of bandwidths to different SUs accordingly. Since the number of mobile devices increases dramatically (the mobile devices are expected to hit 12.1 billion in 2018), it is unrealistic to recruit all mobiles in a specified region. Thus, we present a new architecture with data-driven. In our work, STED samples a relatively smaller scale of SUs to collect the information of SUs' traffic demand and sends to PSP. However, since the number of sample is limited, it is difficult for PSP to learn the precise information of SUs' traffic demands. Hence, we utilize the data-driven approach to deal with uncertainty of the information. Some previous researchers have noticed the issue of distribution uncertainty and tried to employ robust optimization to address this issue. For instance, Lunden et al. [36] propose a non-parametric cyclic correlation in robust computation, which lead the algorithm doesn't require the distribution of users' traffic. Gong et al. in [60] present a model, which consider the distribution uncertainty of received primary signal in spectrum sensing, to determine the robust threshold that can guarantee the false alarm uncertainty. However, there is a lack of study to incorporate data-driven sensing and DP together in spectrum trading system. In our work, we are trying to melt SUs' traffic demand DP into data-driven based spectrum trading. With the proposed scheme, our work effectively preserves each individual SU's traffic demand and maximizes revenue of PSP under data-driven scheme at the same time.

4.7 Conclusion

In this work, we propose a novel spectrum trading architecture consisting of the PSP, the SSP and the STED. Under this architecture, we proposed a novel 3DPP spectrum trading scheme, which jointly employs DP techniques to preserve SUs' demand, and data-driven approach to characterize the uncertainty of SUs' traffic demand. Moreover we mathematically

prove that the data after employing DP mechanism satisfies the data-driven requirements under different ζ-structure probability metrics. Based on the contribution above, we formulate a RA-SP problem to maximize revenue of the PSP. We employ a confidence set by ζ-structure metric to reformulate the problem to a traditional two-stage robust optimization, and developed algorithms. Through simulations, we show the feasible solutions and verify the effectiveness of the proposed 3DPP scheme.

Chapter 5

Optimization for Utility Providers with Differential Privacy of Users' Energy Profile

5.1 Introduction

With the advanced technologies and equipment on computation, communication, automation, controls and sensing, the traditional electric infrastructure is motivated to be modernized into smart grid, which enables two-way communication including electricity and information. Due to the benefits of smart grid, such as efficiency, reliability, and security, the grid modernization has received a lot of attention. For instance, suffering from hurricane and storms, Puerto Rico's electric power authority proposed to rebuild and modernize the power grid through 2027 [61], which includes building distributed microgrids, the use of renewable resources and so on. In 2017, in Illinois, an 18-month investigation study, named as NextGrid [62], is started to define the grid modernization and examine the opportunities and challenges in the future Illinois electric grid and consumers.

Although modernizing the electric grid introduces improvements, smart grid is also facing some challenges, where the most significant one is power outage/interruptions due to the supply and demand mismatch. As we know, a substantial amount of our electricity is generated from the coal, due to its affordable cost and huge coal reserves. In fact, coal-fired power generation takes relatively long time and the gas-fired power has flexible and prompt response to the real-time events in power grids [63]. To prevent power outage and satisfy customers' demand, the utility providers offset fluctuations by using more expensive gas-fired power or pumped-storage electrical power. For example, Siemens operates a number of gas-fired power plants all over the world, which is capable to provide flexible, reliable and efficient power supply.

Moreover, despite the great benefits from two-way flows of electricity and informa-

tion in smart grid, the chances of malicious attacks and risks of privacy leakage increase. Smart metering is a promising solution to forecast and monitor electricity consumption of consumers. The smart meters are installed in each consumer's end (household, company, factory, etc.). The amount of electricity a customer used is measured and saved in an energy profiles, which will be sent to the utility provider at a requested time interval (the frequency can be as few as 1-5 minutes). The provider utility can predict the user's demand accurately, optimize the operation of all distribution resources, and improve the efficiency of the energy network. However, the energy profiles will be a potential target for well-motivated adversaries to compromise the customer's privacy. In this nearly real-time delivery of energy consumption profiles, the attackers can exactly observe the consumer's behavior, by comparing the differences between consumption profiles. For instance, the attackers/eavesdroppers can easily determine whether a consumer is at home by detailed energy consumption data, like TV or washing machine, and further surmise the consumer's house occupancy, meal times, working hours or lifestyle patterns.

There are some research efforts trying to address security and privacy concerns while meeting the requirements in smart grid. For example, Kamto et al. in [64] used encryption to prevent unauthorized access energy profiles. Baumeister in [65] implemented a public key infrastructure in smart grid, which meets most requirements of smart grid, such as scalability and flexibility. In [66], the authors proposed a lightweight Diffe-Hellman authentication mechanism, with Diffe-Hellman key exchange and hash-based authentication technique. Metke et al. in [67] established a secure communication channel if the smart meters are based on trusted computing platform. Nevertheless, the cryptographic solutions can only keep data protected during transmissions, but not for the cases that the adversary compromises the utility providers' servers, or the utility providers themselves are not trustworthy. Under the assumption that the utility provider is semi-honest, i.e., *honest-but-curious* (e.g. [68,69]), in this work, we propose to allow customers to add distributed differential noises to the measured data before the smart meters send it to the utility provider.

Based on the aggregated “noisy" but statistically correct data, we let the utility provider employ data-driven approach to characterize the uncertainty of customers' power demand, match the demand with the supply, and try to minimize the cost of energy generation. We show that the proposed scheme can effectively reduce the power generation cost of the utility provider while preserving the customers' differential privacy in smart grid. Our salient contributions are summarized as follows.

- In our work, we are focusing on integration of data-driven methodology and differential privacy in smart grid scenario. From the user's side, we protect the differential privacy of user's energy profiles. From the utility provider's side, we implement the data-driven methodology to minimize the energy generation cost based on the collected noisy data.

- In order to preserve the individual energy profile privacy, in our scheme, the consumers deploy distributed differential privacy technique before the smart meters sent the energy consumption data to the utility provider. With the distributed differential privacy algorithm, the utility provider can learn the statistic result of consumers' energy demand without compromising each individual consumer's privacy.

- Based on the given set of energy profiles collected by smart meters at consumers' end, the utility provider employs data-driven approach to predicting the total energy demand of consumers in a specific hour. The utility providers can construct references distribution $\mathbb{P}_0$ of customers' demand from the given set of energy profiles collected by smart meters, and ensure that the distance between the ambiguous distribution $\mathbb{P}$ and the reference distribution $\mathbb{P}_0$ is close enough. Due to the added noise and statistical inference, the utility provider cannot predict the future demand accurately. Therefore, in order to meet the fluctuation, gas-fired power plants or pumped-storage electrical power station will be exploited.

- Based on the modeling above, we formulate the cost minimization problem into a risk-averse two-stage stochastic problem (RA-SP). To solve the problem, we utilize

L_1 and L_∞ norms for distance robustness. Our proposed model solves the problem directly from the historical data without assuming/generating the true distribution of consumer's demand. We also verify the effectiveness of the proposed scheme by evaluation performance.

5.2 Network Model

In smart grid, due to the two-way communication of electricity and information, the utility providers are supposed to profile users' demand in order to efficiently balance the supply and demand, at the meanwhile, reduce the cost. In practice, there exists a significant amount of historical data about consumers' demand. With a given set of energy profiles collected by smart meters, at utility provider side, data-driven is applied to forecast the demand. In addition, we assume the utility provider is not trustworthy and the users add the differential noises by themselves. In our work, we make the user perform DDP (distributed differential privacy) algorithm, which is going to be introduced in detail in Subsection 5.2.1 to randomize the true data and send the modified value to the utility provider. As shown in Figure 5.1, with the smart meters, the residential users first process their true energy demand with the DDP algorithm. The utility provider collects the data from the given set of users and applies data-driven model to estimate the energy demands and determine the amount of supply.

In our architecture, the utility provider predicts the future demand from history data of a given set of customers' noisy demand, construct a reference distribution and predict the total energy demand for all customers. Since the reference distribution cannot present 100% unique true demand distribution, the scheduled energy supply may not meet the demand of all customers. Under this scenario, when the supply and demand are not matched, the quick-response efficient gas-fired power plant or pumped-storage electrical power station will be started up. We assume the given set of residential users is $\mathcal{N} = \{1, \cdots, j, \cdots, N\}$ and the real demand for each user is U_j. There are several backup gas-fired power plants/energy

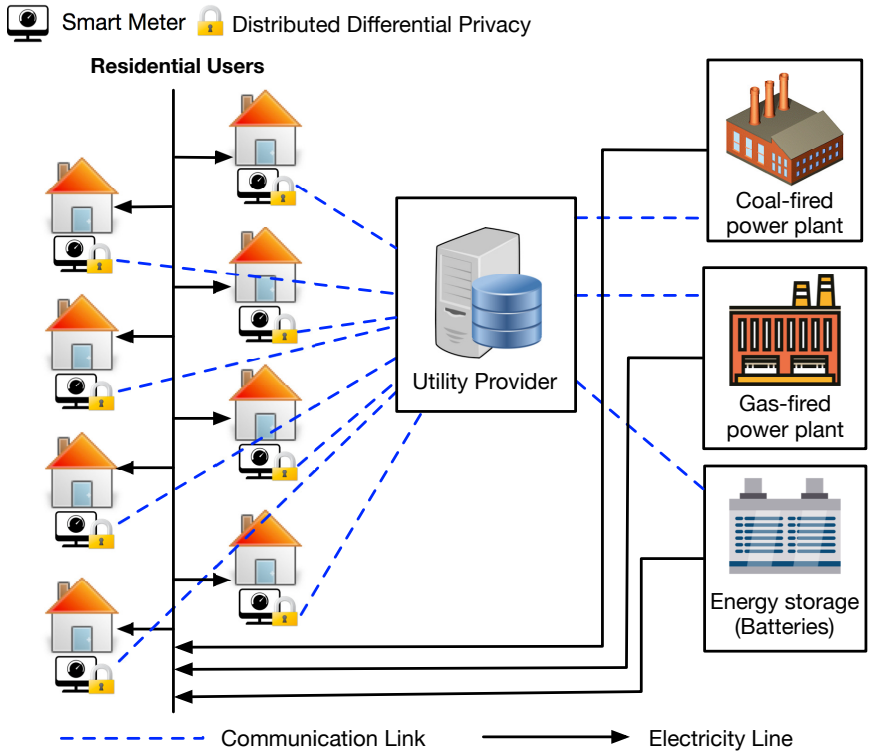

Figure 5.1: Network overview.

storage from a set $\mathcal{M} = \{1, \cdots, i, \cdots, M\}$ controlled and operated by the utility provider. Each has the capacity of c_i unit of electricity power provided to the consumers.

5.2.1 Add Noise with Distributed Differential Privacy Algorithm

The definition of *differential privacy* is first proposed by Dwork [6]. The aim is to exploit the statistical information without disclosure of the data providers' privacy. However, in differential privacy settings, there is a strong assumption that a trustworthy third-party database or data aggregator is required to apply the randomized algorithm on the exact data of data providers. In reality, people may evade to provide data to a survey including sensitive questions. In such a situation that the data providers trust no one even the data collectors but only themselves, secure multi-party computation and homomorphic encryption are suitable to involve in the differential privacy definition. In [5], the authors propose a private stream aggregation algorithm which guarantees distributed differential privacy of

individual user, while the aggregator can only get the statistical results, but not learn any other unintended information from users. We assume $\mathbf{s} = \{s_1, \cdots, s_n\}$ denotes the vector of users' energy demand, function f represents the desired statistics of utility provider and $\mathcal{O}$ indicates the output range of function f. Then, each consumer will employ Distributed Differential Privacy as shown in Sec2.1.2.

In our scenario, since we are focusing on the summation of all the residential users' demand, the sensitivity Δ is supposed to be the maximum energy demand of an individual customer. After achieving the differential privacy, each customer sends the noisy energy demand data encrypted with private key sk_j to the utility provider. With the aggregation of all the encrypted noisy energy demand data, the utility provider is able to sufficiently decrypt the summation of the data [70] with the key sk_0. Therefore, the utility provider is supposed to learn only the summation of users' energy demand and no information from each user. With the DDP algorithm, the utility provider is able to get the summation demand during a time period $d = f(\mathcal{A}(\mathbf{s}))$, which is a noisy version of $f(\mathbf{s})$, at the meanwhile, the differential privacy of each individual customer is guaranteed.

5.2.2 Data-driven Prediction

The traditional two-stage stochastic programming approach under our scenario assumes the distribution of the consumers' energy demand is known. However, in reality, the distribution for the forecasting energy demand is actually uncertain. Instead, only a series of historic consumer's energy profile data are available. In this work, we employ a data-driven approach, i.e., the risk-averse stochastic optimization approach (RA-SP) allowing distribution ambiguity [71], to characterize the uncertainty of forecasting energy demand. In the proposed method, we build the reference distribution from a given set of empirical consumers' data. Since the reference distribution from empirical data might be different from the true distribution, we employ statistical inference and define confidence sets $\mathcal{D}$ corresponding to a given tolerance θ. It allows the distribution ambiguity to be within confident

sets $\mathcal{D}$ with a certain confidence level (e.g., 99%). We employ two norms, L_1 and L_∞ norms, to construct two types of confidence sets D_1 and D_∞. As the number of sampled days is sufficiently large, the reference distribution converges to the true distribution under both two norms. In the following, we describe the two confident sets D_1 and D_∞ as :

$$\begin{aligned}\mathcal{D}_1 &= \left\{ P \in \mathbb{R}_+^K \middle| \|P - P_0\|_1 \leq \theta \right\} \\ &= \left\{ P \in \mathbb{R}_+^K | \sum_{k=1}^{K} \left|p_k - p_k^0\right| \leq \theta \right\} \end{aligned} \tag{5.1}$$

and

$$\begin{aligned}\mathcal{D}_\infty &= \left\{ P \in \mathbb{R}_+^K \middle| \|P - P_0\|_\infty \leq \theta \right\} \\ &= \left\{ P \in \mathbb{R}_+^K | \max_{1 \leq k \leq K} |p_k - p_k^0| \leq \theta \right\}. \end{aligned} \tag{5.2}$$

Under these two norms, the formulated problem can be obtained as a mixed integer linear programming eventually.

For a given set of processed energy profiles data (assuming there are historical data samples), it is easy for us to construct a histogram to fit all the energy profiles data. For example, we can set K intervals to fit the predicted total energy demand of sampled days in each interval to be L_1, L_2, $\cdots$, and L_K with $L = \sum_{k=1}^{K} L_k$. Based on this, we can construct an reference distribution for the uncertain total energy demand of all consumers in particular time period of a day as $p_1^0 = L_1/L$, $p_2^0 = L_2/L$, $\cdots$, and $p_K^0 = L_K/L$. For simplicity, we let $P_0 = p_1^0, p_2^0, \cdots, p_K^0$ represent the corresponding reference distribution.

The two distribution sets under Norm_1 and Norm_∞ are built based on a given confidence level and the amount of available historical data. For instance, β is set to represent the confidence level and $\beta = 98\%$ indicates that the ambiguous distribution P has at least 98% chance in the given set. In (5.1) and (5.2), θ denotes the tolerance value, which is derived from the confident set β and the number of historical data. Intuitively, the more historical data we have, the more "closer" between the reference distribution and true distribution.

From [72], we can explore the precise relationship between the tolerance θ and the number of historical data L. The propositions are shown as follows:

Proposition 1: Supposing there are L number of historical samples, and K intervals, the convergence rate between P and P_0 under L_1 norm is:

$$Pr\{P \in \mathbb{R}_+^K | \|P - P_0\|_1 \leq \theta\} \geq 1 - 2K \exp(-2L\theta/K).$$

Proposition 2: Supposing there are L number of historical samples, and K intervals, the convergence rate between P and P_0 under L_∞ norm is:

$$Pr\{P \in \mathbb{R}_+^K | \|P - P_0\|_\infty \leq \theta\} \geq 1 - 2K \exp(-2L\theta).$$

We can derive the relation between confidence level β and the tolerance θ from above as

$$\theta \text{ for } L_1 \text{ norm} : \theta_1 = \frac{K}{2L} \log \frac{2K}{1-\beta} \quad \text{and} \tag{5.3}$$

$$\theta \text{ for } L_\infty \text{ norm} : \theta_\infty = \frac{1}{2L} \log \frac{2K}{1-\beta}. \tag{5.4}$$

From (5.3) and (5.4), it is easy to observe that, as the size of historical data L increases to ∞, both tolerance θ_1 and θ_∞ decrease to 0. Therefore, the confidence sets $\mathcal{D}_1$ and $\mathcal{D}_\infty$ become singleton, and the corresponding risk-averse two stage stochastic problem becomes the traditional two-stage stochastic problem.

5.2.3 Cost Minimization Problem Formulation

The collected data with DDP from smart meters is aggregated by the utility provider. Because the distribution of the customers' demand is uncertain, in order to efficiently balance the supply and demand, the utility provider employs data-driven approach to forecasting the customers' future demand. As the reference distribution constructed from the collected data cannot present the unique true distribution of the customers' demand. In order to match the supply and demand, the quick-response gas-fired power plants or pumped-storage electrical

power stations are used. At the same time, on the utility provider side, the cost is supposed to be minimized. Consequently, the cost minimization problem for utility provider can be formulated as

$$\min_{x,y} \sum_{i}^{N} F_i y_i + \mathbb{E}_{\mathbb{P}}\Big[\sum_{i}^{M} T_i x_i(\xi)\Big], \tag{5.5}$$

s.t.:

$$\sum_{i}^{M} x_i(\xi) \leq c_i y_i \quad \forall i, \tag{5.6}$$

$$\sum_{i}^{M} x_i(\xi) = \sum_{j} U_j - d(\xi), \text{ and} \tag{5.7}$$

$$x_i(\xi) \geq 0, y_i \in \{0,1\} \quad \forall i, j. \tag{5.8}$$

In the formulation, (5.6) indicates the energy generated from gas-fired power plant i should not exceed its capacity and (5.7) indicates the total number of energy generated from all gas-fired power plants is the gap between overall real demand and uncertain predicted demand from energy utility. The opening price for each gas-fired power plant is represented by F_i, y_i is a binary variable indicating if gas-fired power plant i is open, T_i denotes the purchase price of each unit from power plant i, x_i is the energy generated from the power plant i, c_i expresses the capacity of each power plant and U_j means real demand from each consumer j in the particular time period.

Since we add noise in the processed energy profile, the distribution of real demand is ambiguous. Therefore, we construct the confident set $\mathcal{D}$, and let $P \in \mathcal{D}$ so as to minimize the total cost under the worst-case distribution realization in $\mathcal{D}$. The detailed formulation is described as

$$\min_{y} \sum_{i}^{M} F_i y_i + \max_{p_k} \sum_{k}^{K} p_k \min_{x} \sum_{i}^{M} T_i x_i(\xi_k), \tag{5.9}$$

s.t.: (5.6) – (5.8),

$$\sum_{k=1}^{K} p_k = 1, \text{ and} \tag{5.10}$$

$$P \in \mathcal{D}. \tag{5.11}$$

5.2.4 Solution to the Optimization Problem

The Benders' decomposition algorithm [73] is exploited to solve the problem into global optimality. Since for each scenario ξ_k, the second-stage optimization problem $\min_x \sum_i^M T_i x_i(\xi^k)$ of (5.9) is independent of ξ_i for $i \neq k$. Consequently, the minimization operation can be put before the summation, i.e, the objective function (5.9) can be written as

$$\min_y \sum_i^M F_i y_i + \max_{p_k} \min_x \sum_{k=1}^{K} p_k \sum_i^M T_i x_i(\xi_k), \tag{5.12}$$

$$\text{s.t.:} \qquad (5.6) - (5.8), (5.10), (5.11).$$

We can calculate the second-stage minimization problem by solving its dual. The dual subproblem and dual variables λ, v associated with constraints are given by

$$\max_{\lambda,v} \sum_{k=1}^{K} \left[\lambda_k \left(U - d(\xi_k) \right) - \sum_i^M c_i y_i v_k^i \right], \tag{5.13}$$

$$\text{s.t.:} \qquad \lambda_k - v_k^i \leq p_k T_i, \forall i, k, \text{ and} \tag{5.14}$$

$$v_k^i \geq 0, \forall i, k. \tag{5.15}$$

The dual variables corresponding to scenario k for constraints (5.6)-(5.8) are v_k^i and λ_k, respectively. Because of the duality property, the optimal objective of the primal problem is equivalent to the dual problem. It is obvious that the maximization operation in the primal formulation can be combined with the dual second-stage problem. The second-stage max-min problem can be obtained as

$$\psi(y) = \max_{p_k} \min_x \sum_{k=1}^{K} p_k \sum_i^M T_i x_i(\xi_k)$$

$$= \max_{p_k,\lambda,v} \sum_{k=1}^{K} \left[\lambda_k \left(U - d(\xi_k) \right) - \sum_{i}^{M} c_i y_i \upsilon_k^i \right], \tag{5.16}$$

$$\text{s.t.:} \qquad \lambda_k - \upsilon_k^i \leq p_k T_i, \forall i, k, \tag{5.17}$$

$$\upsilon_k^i \geq 0, \forall i, k, \text{ and} \tag{5.18}$$

$$\sum_{k=1}^{K} p_k = 1, \qquad P \in D. \tag{5.19}$$

Under the L_∞ norm case, the constraint (5.19) represents

$$\max_{1 \leq k \leq K} |p_k - p_k^0| \leq \theta, \tag{5.20}$$

which is equal to

$$|p_k - p_k^0| \leq \theta, \forall k. \tag{5.21}$$

Under the L_1 norm case, the constraint (5.19) represents

$$\sum_{k=1}^{K} |p_k - p_k^0| \leq \theta. \tag{5.22}$$

We denote α as the second-stage worst case energy cost. Then by applying feasibility cut and optimality cut iteratively, we can solve the master problem which is reformulated as

$$\min_{y \in \{0,1\}} \sum_{i}^{N} F_i y_i + \alpha$$

$$\text{s.t.: Feasibility cuts,}$$

$$\text{Optimality cuts.}$$

- *Feasibility Cuts*: We use the L-shaped method to generate feasibility cuts. We formulate the feasibility check problem as follows to check constraints (5.6) and (5.7):

$$\min_{\gamma,x} \sum_{k=1}^{K} \left(\sum_{i=1}^{M} \gamma_k^{1i} + \gamma_k^2 + \gamma_k^3 \right), \tag{5.23}$$

$$\text{s.t.} \qquad \gamma_k^{1i} - X_i(\xi_k) \geq -c_i y_i, \forall i, k, \tag{5.24}$$

$$\gamma_k^2 - \sum_{i=1}^{M} X_i(\xi_k) \geq -(U - d(\xi_k)), \forall k, \tag{5.25}$$

$$\gamma_k^3 - \sum_{i=1}^{M} X_i(\xi_k) \geq (U - d(\xi_k)), \forall k, \text{ and} \tag{5.26}$$

$$X_i(\xi_k) \geq 0, \gamma_k^{1i}, \gamma_k^2, \gamma_k^3 \geq 0, \forall i, k. \tag{5.27}$$

Its dual problem can be obtained as

$$\omega(y) = \tag{5.28}$$

$$\max_{\hat{\lambda},\hat{\mu},\hat{v}} \sum_{k=1}^{K} \left[-\hat{\lambda}_k \left(U - d(\xi_k)\right) + \hat{\mu}_k \left(U - d(\xi_k)\right) - \sum_{i=1}^{M} \hat{v}_k^i c_i y_i \right] \tag{5.29}$$

$$\text{s.t.} \qquad -\hat{\lambda}_k + \hat{\mu}_k + \hat{v}_k^i \leq 1, \forall i, k, \tag{5.30}$$

$$\hat{\lambda}_k, \hat{\mu}_k, \hat{v}_k^i \in [0, 1], \forall i, k, \tag{5.31}$$

where dual variables $\hat{v}_k^i$, $\hat{\lambda}_k$ and $\hat{\mu}_k$ correspond to the kth scenario for constraints (5.24), (5.25) and (5.26), respectively. Therefore, the feasibility check is performed as follow steps:

1) If $\omega(y) = 0$, the first stage solution is feasible.

2) If $\omega(y) \geq 0$, a feasible cut is generated in the following form:

$$\sum_{k=1}^{K} \left[-\hat{\lambda}_k \left(U - d(\xi_k)\right) + \hat{\mu}_k \left(U - d(\xi_k)\right) - \sum_{i=1}^{M} \hat{v}_k^i c_i y_i \right] \leq 0. \tag{5.32}$$

- *Optimality Cuts*: At each iteration, we get y and α after solving the master problem. Then we substitute y into the subproblem and obtain $\psi(y)$. If $\psi(y) \leq \alpha$, we claim we find the optimal solution. If not, which means $\psi(y) > \alpha$, we generate optimal cut in the following form and add it into the master problem:

$$\sum_{k=1}^{K} \left[\lambda_k \left(U - d(\xi_k)\right) - \sum_{i}^{M} c_i y_i v_k^i \right] \leq \alpha. \tag{5.33}$$

Finally, the optimality cuts and feasibility cuts ensure the Benders' decomposition algorithm converges to global optimality.

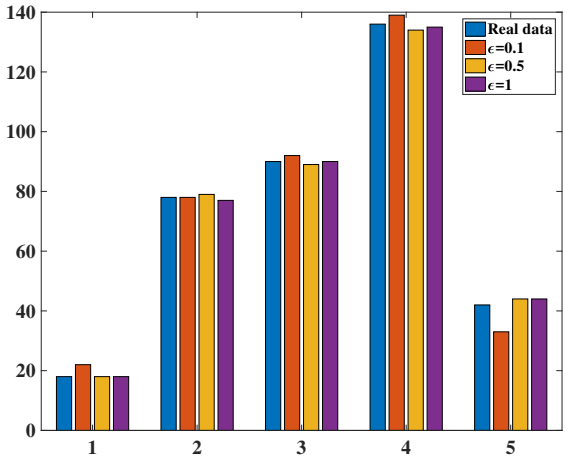

Figure 5.2: Distribution of consumers' energy demand under different ϵ.

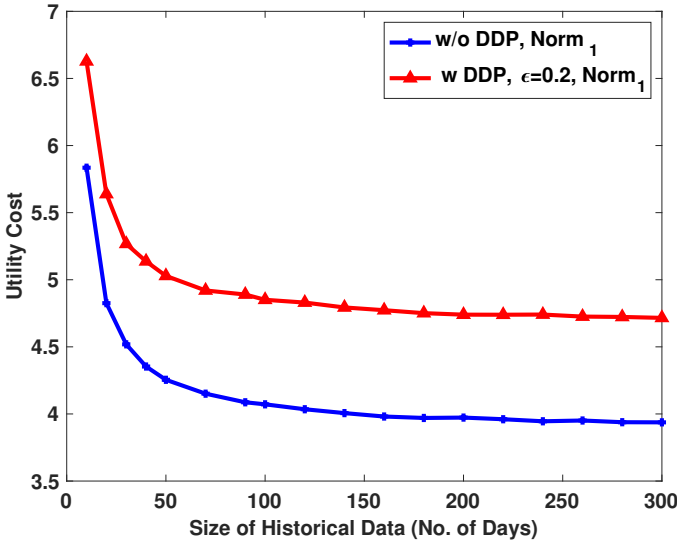

Figure 5.3: Total cost under L_1 norm

5.3 Performance Evaluation

In this section, we evaluate our proposed model and associated algorithms. The evaluation is accomplished in a computer equipped with Intel Core i7 CPU of 2.7GHz. Due to the nondisclosure agreement, all results are computed from the simulated data that are generated according to the real data analysis. The proposed algorithms can be directly applied to real data without modification. The utility provider processes 10,000 consumers energy cost per hour (from 8 pm to 9 pm) for 300 days. The consumers implement DDP algorithm to their energy profiles. We assume the total consumers' energy consumption is 6×10^4. In our model, there are three gas-fired power plants, each capacity is 4×10^4; 5×10^4 and 1×10^4, accordingly. In addition, the open cost for each gas-fired power plant is 1, and the unit wholesale price for energy is 1 per unit.

We set the confidence level β to 80% and study the data-driven algorithm without DDP. The results are shown in Fig. 5.2−Fig. 5.4. From Fig. 5.2, we can observe that the distribution of consumers' energy demand is very close after integrating distributed differential privacy. We notice that as ϵ get smaller, the privacy is higher, therefore, the difference between the distributions is higher too. In Fig. 5.3 and Fig. 5.4, we obtain the performance of the energy generation cost under different forms. It is shown that under both L_1 norm and L_∞ norm, the cost is lower as we have more historical data. It means that with more historical data processed by utility provider, the distribution is more accurate. It can also be observed that the utility cost under DDP algorithm is worse than the performance without preservation privacy, which represents the trade-off between the privacy and utility.

5.4 Conclusion

In our work, we focus on integration of data-driven methodology and differential privacy in smart grid scenario and propose a novel scheme that not only minimizes the cost for utility providers but also preserves the DDP of users' energy profile via differential privacy.

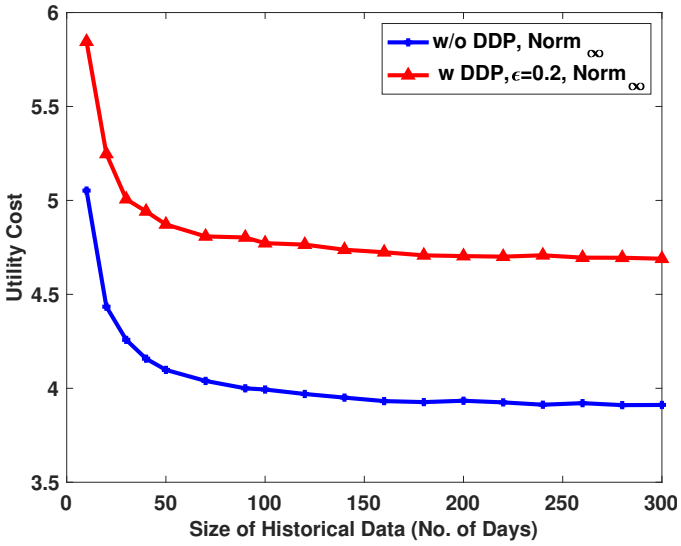

Figure 5.4: Total cost under L_∞ norm

The data-driven approach is exploited to estimate the users' demands and formulate the cost minimization based on the collected noisy data. We also develop algorithms for feasible solutions, and demonstrate the trade-off between privacy and utility of the proposed scheme through simulations results.

Chapter 6

Caching with Users' Local Differential Privacy in Information-Centric Networks

6.1 Introduction

As the rapid increasing of content demands in the Internet, new information-centric networking (ICN) design is motivated to be developed in the future Internet for improved delivery efficiency, content scalability and availability [74–76]. In addition, ICN architectures are based on named content, which is radically different from the traditional host-centric paradigm based on named hosts [74]. In this new ICN architecture, with the deployment of in-network storage for caching in the access points (AP), it is efficient to offload the tremendous increasing amount of content. In 2017, Cisco highlights that the video traffic has already reached 73 percent of all the Internet traffic in 2016, and it is estimated to be increased to 82 percent by 2021 [77]. Inspired by the fact of the speedy growth of the demand for video contents, build-in caching features are supposed to be applied widely in the ICN.

Since the content provider (CP) aims to provide high quality of service (QoS) to the users, the storage for caching in APs plays an important role in reducing the network congestion and backhaul load. As the cache-enabled APs such as base stations are required to cooperate with the CP, it is necessary to find an approach to offering the economic incentives for the contributions and efficiently allocating the resources. As a result, the CP is able to cache the popular data objects with the cooperation of the cache-enabled APs by offering appreciable economic incentives. For example, in [75], the authors exploit the auction theory to design the optimal allocation with jointly leasing the cache storage and bandwidth of APs. In [78], the proposed scheme focuses on optimal virtual resource allocation with integrating device-to-device communication in the ICN. In [79], the non-cooperative game is applied to

formulate a pricing strategy and caching policy. However, in these works, they all use the Zipf discrete distribution [80] to represent the content popularity in Internet. As the universal Zipf distribution may not perfectly capture the statistical features of content popularity in various geographical locations in ICN, in our work, we employ data-driven methodology to predict the content popularity from the collected data of local CP users without premise on the content popularity distribution.

While it is beneficial to ICN users with high QoS, it may compromise the users' privacy. To predict the content popularity, the CP aggregates the preferred content information from certain users. However, this aggregation process may elevate risks of privacy leakage. As the user's content preferences may include some sensitive information, these kind of sensitive personal information could be sold as a commodity for commercial uses. For example, because of the disclosure of the private content preference data, users may receive a plenty of spam or fraud emails or phone calls. Therefore, it is necessary to pay attention on protecting on users' private content preferences. For instance, in [76], the authors propose a tag forgery based privacy-enhancing technology to protect the users' interests and preferences in social-tagging systems. In [81], the authors design a tag suppression scheme based on data perturbation to protect end-user privacy in collaborative tagging services.

In our work, in order to address those issues above, we propose a scheme that the CP offloads popular contents into several storage for caching of APs according to the noisy content preference data from users. Therefore, the users' privacy is preserved and the problem of high backhaul load is resolved. Briefly, the CP exploits the data-driven methodology to predict the content popularity distribution according to the collected noisy content preference data from the users and stimulates the APs with economic incentives to lease their storage for caching popular contents. Consequently, we formulates a revenue maximization problem based on the description above and demonstrates that the CP revenue can be effectively optimized, while preserving the customers' local differential privacy in the ICN. Our salient contributions are summarized as follows:

- In our work, we preserve the privacy of users' preference information by locally adding differential noises on the users' side. In addition, data-driven methodology is employed to forecast the content popularity for optimize the revenue maximization problem on the CP side. However, true preference content information of each individual user is not able to be obtained by CP or attackers like eavesdropper.

- With the assumption that the CP is semi-honest, in order to protect each individual user's content preference information, optimized local hashing (OLH) protocol is exploited. Therefore, the CP is able to estimate the frequency distribution of different contents from the users' noisy content preference information. At the meanwhile, the true individual user's content preference information is not able to be leaked out.

- In the ICN, the Zipfs law is widely applied as a probabilistic model to characterize the content popularity. In our work, we employ data-driven approach to predicting the content popularity of a group of users without assumption of the distribution. We assume the CP constructs the reference content popularity probability $\mathbb{P}_0$ according to the noisy users' content preferences, stimulates the cache-enabled APs to cooperate in the ICN to store popular content and formulates the revenue maximization problem with the constraint of characteristic of uncertainty of content popularity with distance between the ambiguous distribution $\mathbb{P}$ and the reference content popularity probability $\mathbb{P}_0$.

- The formulated revenue maximization problem can be represented into a risk-averse two-stage stochastic problem (RA-SP). The Benders' decomposition is algorithm is applied to solve the proposed problem to global optimality based on L_1 and L_∞ norms for distance robustness. We also conduct simulations to verify the effectiveness of the proposed scheme.

6.2 Network Model and Preliminaries

6.2.1 System Description

In our work, as shown in Figure 6.1, we assume the content provider (CP), in the information-centric network (ICN), collects users' content preferences information with local differential noise, forecasts the content popularity by data-driven methodology, leases several storage for caching of the access points (APs) and offloads the popular contents in advance into the cache. Therefore, the heavy back haul load and congestion problem can be reduced. Additionally, the users apply the local differential privacy (LDP) protocols to add noise individually on their content preferences and send the modified value to the CP.

In our scheme, we assume the set of users is $\mathcal{U} = \{1, \cdots, u, \cdots, U\}$, the file is represented as f with size s_f and the real content preference of each user is r_u that is in the domain with size of F. There are several cache-enabled APs from a set $\mathcal{M} = \{1, \cdots, m, \cdots, M\}$ cooperated with the CP to provide high QoS. Each cache of the AP has the capacity of c_m unit and the price to lease each cache is k_m. With the LDP protocol, the users add noise locally to their content preference r_u, which is shown in 6.3.1 in detail. The CP constructs the reference content popularity probability $\mathbb{P}_0$ based on noisy content preference results and predicts the true popularity by data-driven approach. Hence, the storage for caching in APs is selected to lease by CP. Because of the uncertainty of reference distribution, the revenue maximization problem is formulated to determine the set of cache to be leased, which is illustrated in Sec 6.3.3. Moreover, the Benders' decomposition is deployed to solve the proposed maximization problem.

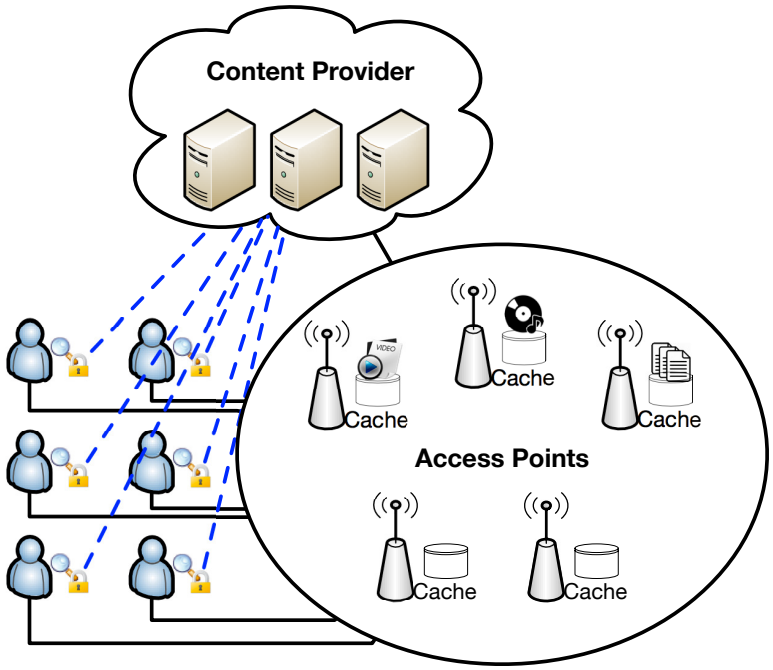

Figure 6.1: System description.

6.3 Data-Driven Caching Revenue Maximization Problem with Local Differential Privacy

6.3.1 Protecting Private Content Preference with Local Differential Privacy

To perform a LDP mechanism, it contains several steps. First, the true data is encoded locally into a vector or a number. Next, the encoded data is randomized by a specific function. At last, the processed data will be sent to the data aggregator or database. Among the three steps, the combination of the first two steps is the randomized algorithm $\mathcal{A}$ in the definition, which is finished locally and supposed to satisfy ϵ-LDP. In [9], the authors have introduced an optimized LDP protocol, named optimal local hashing (OLH), which can offer higher accuracy of frequency estimation with lower communication cost.

In our work, we exploits OLH protocol to randomize the response of users' content preferences. The detail of OLH protocol can be found at Sec 2.1.3.

Algorithm 6.1 Algorithm for Obfuscation Strategy

1: **Input:** Survey data of user preference i.i.d drawn from the true distribution. The confident level of D is η.
2: **Output:** Objective value of the problem (6.1).
3: Obtain the reference distribution $\mathbb{P}_0$ and tolerance θ based on the historical data.
4: **if** The reference distribution and true distribution are under Kantorovich metric or Fortet-Mourier metric **then**
5: Reformulate the problem to (6.12) - (6.14)
6: Feasibility check master problem of (6.12)
7: **if** Infeasible **then**
8: Generate feasible cut for master problem
9: go to line 6
10: **end if**
11: Feasibility check the subproblem of (6.12)
12: **if** Infeasible **then**
13: Generate optimal cut for subproblem
14: go to line 6
15: **end if**
16: Stop and output solution
17: **else**
18: Reformulate the problem to (6.15) - (6.17) under Uniform metric
19: Solve the problem under bender decomposition algorithm, same as line 6 to line 16.
20: Output the solution.
21: **end if**

6.3.2 Data-driven Analysis of Content Popularity

Most works in the ICN assume that the distribution of content popularity is known as Zipf distribution. However, practically, it characterizes the the statistical features in various geographical locations. Moreover, only historical data or real content preferences of users can be obtained by the CP to construct the reference distribution of content popularity. Therefore, in our work, we employ data-driven risk-averse stochastic optimization approach (RA-SP) to making a decision to lease cache-enabled APs under the uncertainty of predicting the content popularity.

We use a distance measurement proposed in [10] to quantify the distance between two distributions. Specifically, a predefined distance measure $d(\mathbb{P}_i^0, \mathbb{P}_i)$ is constructed on confident set D, where $\mathbb{P}_i^0$ is the reference distribution estimated from historical data, and $\mathbb{P}_i$ is the ambiguous distribution of users' content preferences distribution. The details can be found at Sec 2.2.

6.3.3 Caching Revenue Maximization Problem with Local Privacy Preservation

As we describe before, the CP collects user's noisy content preferences with LDP and aggregates the frequency estimation of each content. Consequently, the CP is able to get the noisy content popularity represented as $r_u(f)$. In our work, according to the noisy content popularity $r_u(f)$, we assume there are F popular files in the set $\mathcal{F} = \{1, \cdots, f, \cdots, F\}$ selected to store in the cache-enabled APs, each of which has the capacity c_m. The binary parameter y_m is used to represent if an AP is leased by the CP with the price k_m. We denote the size of a file stored in a cache as s_{fm}. We represent the profit per unit size of backhaul load reduction as α. Hence, the total expected revenue of backhaul load reduction is $\alpha \sum_f^F \sum_m^M s_{fm}$. We randomly sample select a group of users of the CP and get the total download size of files. Since the uncertainty of the distribution of the content popularity, the total download size distribution can be denoted as $d(\xi)$. In addition, the profit of serving users per unit size of file is ϕ. In order to maximize revenue, the CP employs data-driven method to predict the real content popularity and selects cache-enabled APs from a given group. Because of contribution of the APs, the bachhual load is reduced. Therefore, the revenue maximization problem for CP can be formulated as

$$\max_{y} \sum_{m}^{M} -k_m y_m + \alpha \sum_{f}^{F} \sum_{m}^{M} s_{fm} + \mathbb{E}_{\mathbb{P}}[\phi d(\xi)], \tag{6.1}$$

$$\text{s.t.:} \quad \sum_{f}^{F} s_{fm} \leq c_m y_m, \forall m, \text{ and} \tag{6.2}$$

$$y_m \in \{0, 1\}, \forall m. \tag{6.3}$$

In the formulation, (6.2) indicates the files store in one cache m should not exceed the capacity c_m of the cache and (6.3) indicates whether the cache m is leased by the CP. Since we add noise in the processed energy profile, the distribution of real demand is ambiguous.

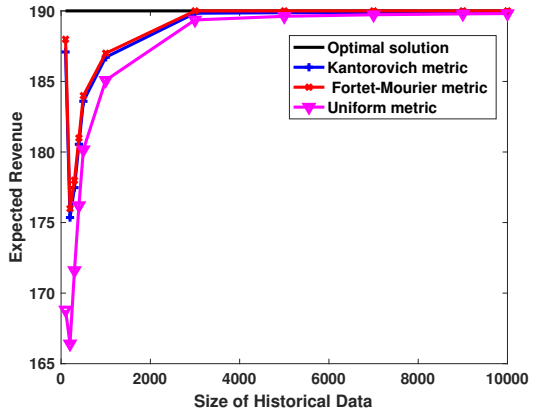

Figure 6.2: Expected revenue without users' privacy preservation.

Therefore, we construct the confident set $\mathcal{D}$, and let $P \in \mathcal{D}$ so as to minimize the total cost under the worst-case distribution realization in $\mathcal{D}$. The detailed formulation is described as

$$\max_{y} \sum_{m}^{M} -k_m y_m + \alpha \sum_{f}^{F} \sum_{m}^{M} s_{fm} + \min_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}[\phi d(\xi)], \tag{6.4}$$

$$\text{s.t.} \qquad (6.2), (6.3).$$

6.3.4 Solution to Caching Optimization under Distribution Uncertainty

From subsection 6.3.2, we explore how to solve (6.4). We assume the sample space is $\Omega = \{\xi^1, \xi^2, \cdots, \xi^N\}$. The formulation can be simplified as :

$$\max_{y} \sum_{m}^{M} -k_m y_m + \alpha \sum_{f}^{F} \sum_{m}^{M} s_{fm} + \min_{p_i} \sum_{i=1}^{N} p_i\big(\phi d(\xi^i)\big), \tag{6.5}$$

$$\text{s.t.} \qquad (6.2), (6.3),$$

$$\sum_{n=1}^{N} p_i = 1, \text{ and} \tag{6.6}$$

$$\max_{h_i} \sum_{i=1}^{N} h_i p_i^0 - \sum_{i=1}^{N} h_i p_i \le \theta, \forall h_i : ||h||_\zeta \le 1, \tag{6.7}$$

where $||h||_\zeta$ is defined according to different metrics. For the Kantorovich metric, $|h_i - h_j| \leq \rho(\xi^i, \xi^j)$. For the Fortet-Mourier metric, $|h_i - h_j| \leq \rho(i,j)\max\{1, \rho(\xi^i, a)^{p-1}, \rho(\xi^j, a)^{p-1}\}$. The constraints (6.6)-(6.7) can be summarized as $\sum_{i=1}^{N} a_{ij}h_i \leq b_j, j = 1, \cdots, J$. To reformulate the constraints, we consider the following problem:

$$\min_{h_i} \quad \sum_{i=1}^{N} h_i p_i^0 - \sum_{i=1}^{N} h_i p_i, \tag{6.8}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} a_{ij}h_i \leq b_j, j = 1, \cdots, J. \tag{6.9}$$

The dual problem can be formulated as

$$\min_{u} \quad \sum_{j=1}^{J} b_j u_j, \text{ and} \tag{6.10}$$

$$\text{s.t.} \quad \sum_{j=1}^{J} a_{ij}u_j \geq p_i^0 - p_i, \forall i = 1, \cdots, N, \tag{6.11}$$

where u is the dual variable. Accordingly, the problem can be reformulated as follows under Kantorovich metric and Fortet-Mourier metric:

$$\max_{y} \sum_{m}^{M} -k_m y_m + \alpha \sum_{f}^{F}\sum_{m}^{M} s_{fm} + \min_{p_i} \sum_{i=1}^{N} p_i\big(\phi d(\xi^i)\big), \tag{6.12}$$

$$\text{s.t.} \quad (6.2), (6.3),$$

$$\sum_{n=1}^{N} p_i = 1, \sum_{j=1}^{J} b_j u_j \leq \theta, \text{ and} \tag{6.13}$$

$$\sum_{j=1}^{J} a_{ij}u_j \geq p_i^0 - p_i, \forall i = 1, \cdots, N. \tag{6.14}$$

For the Uniform metric, we can have the reformulation from the Uniform metric definition:

$$\max_{y} \sum_{m}^{M} -k_m y_m + \alpha \sum_{f}^{F}\sum_{m}^{M} s_{fm} + \min_{p_i} \sum_{i=1}^{N} p_i\big(\phi d(\xi^i)\big), \tag{6.15}$$

s.t. (6.2), (6.3),

$$\sum_{i=1}^{N} p_i = 1, \text{ and} \tag{6.16}$$

$$\left|\left(p_i^0 - p_i\right)\right| \leq \theta, \forall i. \tag{6.17}$$

After reformulating the problem, we can solve the formulation (6.12) - (6.14) and (6.15) - (6.17) through Benders' decomposition algorithm. The detailed algorithm is shown in Algorithm 6.1.

6.4 Performance Evaluation

In our simulation, we assume the CP provides service to 10,000 users and take a survey on the content preference from the selected users. The users implement local differential privacy protocol, then send the noisy preference result to the CP. To be specific, the sample users choose interested files from 10 candidates, add noise and send back to the CP. The CP processes the results and estimates the maximum revenue. We assume there are three cache-enabled APs in our system, each capacity is 3 units, 4 units and 5 units, accordingly. To simplify simulation, each file is 4 units in our network.

We set the confidence level β to 99% and study the performance with and without integrating local differential privacy. In Fig. 6.2, we obtain the performance of the expected revenue under different metrics. It is shown that no matter under which metric, the expected revenue is higher and closer to optimal revenue as we have more data. It means that with more data processed by the CP, the distribution is more accurate. Fig.6.3 − Fig.6.5 show the comparison under different local differential privacy levels. We can observe that the expected revenue after adding noise is worse than the performance without preserving privacy, which represents the trade-off between the privacy and utility. Moreover, it is observed that when ϵ increases from 0.5 to 1, the expected revenue increases under the same size of historical data, and closer to the performance without adding noise. The reason is that ϵ presents

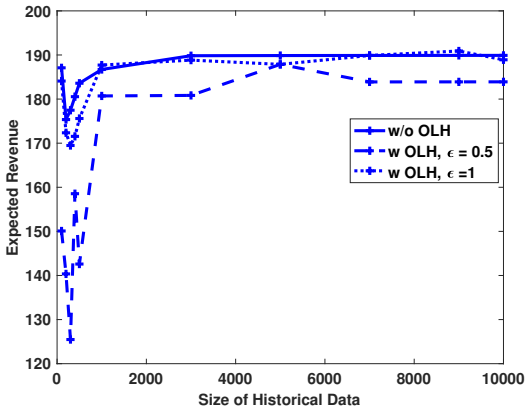

Figure 6.3: Performance under Kantorovich metric .

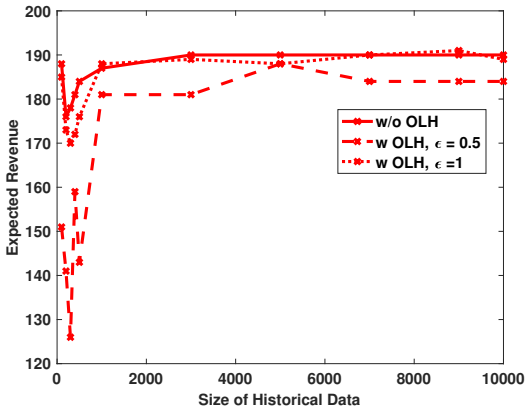

Figure 6.4: Performance under Fortet-Mourier metric.

privacy level in differential privacy. When ϵ is smaller, it means the privacy level is higher, and the users would add more noise in the submitted data.

6.5 Conclusion

In our work, we propose a scheme to predict the content popularity based on selected users' locally differentially private content preference data in information-centric networks and formulate a revenue maximization problem. Because of the uncertainty of the content popularity distribution, data-driven methodology is employed to formulate the problem based on the collected noisy content preference data. In addition, we develop an algorithm for obfuscation strategy to feasibly solve the proposed problem. We conduct simulations to

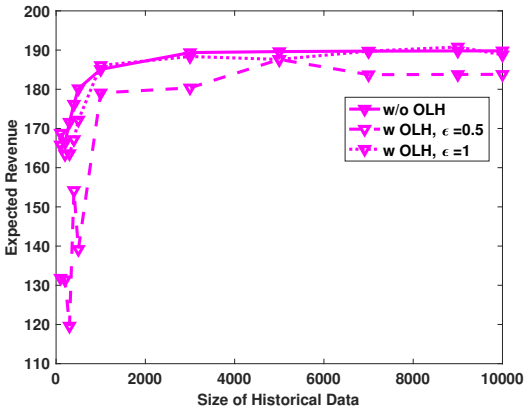

Figure 6.5: Performance under Uniform metric.

show the effectiveness of the proposed scheme and illustrate the trade-off between privacy and utility.

Chapter 7

Data-Driven Small Cell Placement Optimization with Users' Differential Privacy for Wireless NGNs

In the coming fifth generation (5G) or beyond 5G next generation networks (NGNs), the small cell deployment is a promising solution to meet the ever increasing demands of mobile devices, and the proliferation of wireless services. The low power base station (BS), such as femtocell BS, is a cost-effective and environmental friendly substitution for the power-hungry macrocell BS. One potentially effective way to deploy those small cells is to use two-tier NGN architecture, where the first-tier carrier can authorize the second-tier carrier's access to users' transmission information database (e.g., uplink/downlink service demands), and thereafter the second-tier carrier can decide how to place small cell BSs according to the mobile users' requirements locally. However, the second-tier carriers/operators for small cell placement may not be trustworthy, and the NGN users' data privacy might be compromised. To address this issue, we integrate differential privacy (DP) preserving techniques into data-driven optimization, and propose a novel scheme that not only preserves the privacy of NGN users' transmission information, but also maximizes the revenue of small cell deployment. Briefly, differential private noises are intentionally added into the users' transmission information database. Based on queries, the second-tier carrier can aggregate a given set of users' differentially private historical data, estimate the users,Äô demands, and formulate the data-driven revenue maximization problem. Given the stochastic programming optimization formulation, we develop feasible solutions and conduct extensive simulations with real-world transmission datasets (i.e., transmission data collected hourly from 3072 4G eNBs deployed in several southern cities of China in 2015) to verify the effectiveness of the proposed scheme.

7.1 Introduction

Over the last decades, the world has witnessed the explosive growth of mobile wireless communication, fueled by the popularity of smartphones and tablets. By 2020, it is expected that almost 50 billion wireless devices will be connected [82,83], and each mobile users would download 1 terabyte of data annually [84]. However, the existing wireless communication technologies cannot handle the thousand-fold increase in total mobile broadband data in the future [85]. Therefore, the NGN wireless communication technologies have been proposed to support the demands for a higher data rate, a larger capacity, and a lower latency, compared to 4G wireless technologies, such as LTE. It is predicted that NGNs will rely much more on the small cell, such as femtocell with low-power, short range access points to deliver data to the users [86, 87].

To address the aforementioned challenges, the multi-tier architecture is one of the promising solutions for NGNs. To be specific, in a particular region with fixed number of users, to meet the high data rate demand in NGNs, one macrocell BS cannot satisfy all the users' traffic demands. Traditionally, the first-tier carrier will build another macrocell BSs to ensure that more than one macrocell BS will cover this region to achieve the high data rate demand. However, maintaining a macrocell BS consumes a large amount of energy and money. A multi-tier architecture consists of macrocell BS, different types of licensed small cells to serve users with different quality-of-service (QoS) requirement in an energy-efficient manner. Instead of building another macrocell BS, the second-tier operator will build many small cell BSs, such as microcell, and femtocell users, which can collaborate with macrocell BS in a multi-tier architecture, to support the capacity and increase the spectral efficiency. An example of a tier-2 carrier in the US is US Cellular, who has an agreement in place with Sprint for voice and data coverage. The advantage of small cell BSs is that they are more energy-efficient choice. (One microcell BS consumes about 4.4 times less energy than a macrocell BS [88]), which can reduce the inter-cell interference and improve the spatial

reuse if necessary [89].

7.1.1 Related Work

In the multi-tier cellular wireless NGNs, the first-tier carrier will authorize second-tier carrier to access the data transmission profile of the users, and the second-tier carrier will decide where and how many microcell BSs to build in a particular region. Therefore, the data transmission profile will be a potential target for the well-motivated adversaries if the second carriers are compromised. When the second-tier carrier is untrusted, the attackers/eavesdroppers can easily know the amount of data transferred in a particular region. Therefore, it can estimate the market scale of a hotel, or the day routine of a specific user.

There are many research works that focus on addressing the security and privacy concerns in the small cell [90]. For example, unlinkable temporary identifiers (TMSIs [91] and GUTIs [92]) are employed in 4G LTE standards, to prevent the air interface attacks between the mobile device (Users equipment) and the femtocell (HomeNodeB). The large scale deployment of small small cell BSs also renders the exposure of the core NGN's point of entrance to the public internet, which could make the system under internet-based attack, such as Denial of Service (DoS) or impersonation attack. The most popular solutions for such attacks are network analysis [93, 94] and client puzzles [95]. Guri et al [96] tested and analyzed the anonymous attack on a small cellular network under the current 911 infrastructure, to measure the severity of the attack impact.

7.1.2 Our Contribution

Nevertheless, none of above work consider that the small cell BSs themselves may be untrustworthy. Under the assumption that the small cell BS may be compromised by the adversaries, in this work, we propose a novel scheme, which allows the first-tier carrier could add noise to the user's database under differential privacy standard. After that, the

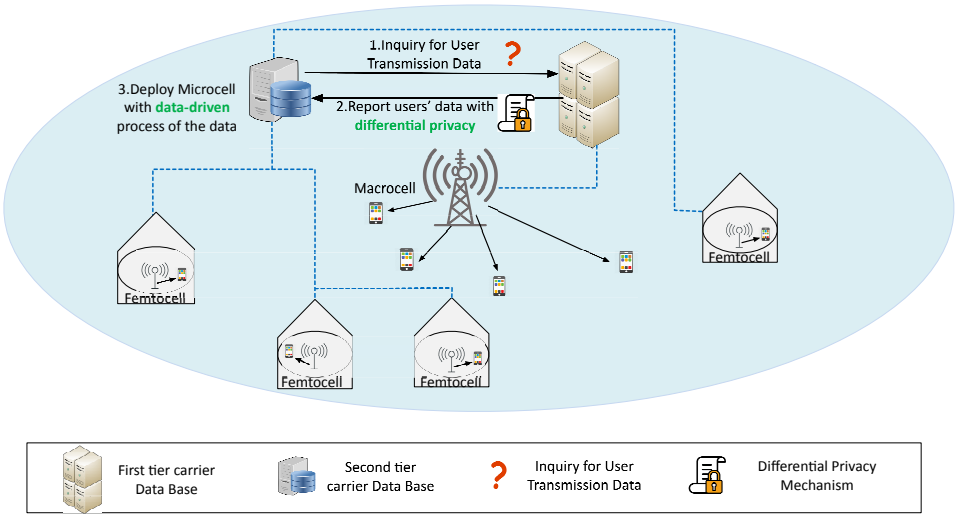

Figure 7.1: Illustration for small cell deployment in NGNs.

first-tier send the data to the small cell BS of second-tier carrier. Based on the "noisy" but statistically correct data transmission amount, we let the small cell BSs employ data-driven approach to characterize the uncertainty of users' transmission data requirement, match the demand with the supply, and try to maximize the total revenue of whole system. It is proved that this novel strategy increases the financial profit of the whole system while preserving the users' communication data differential privacy in NGNs small cell network. Our contributions are summarized as follows,

- In our work, we focus on the integration of differential privacy and data-driven stochastic programming methodology in the small cell deployment. From the first-tier carrier side, we protect the privacy of users' transmission data; from the second-tier carrier/small cell BSs' side, we implement the data-driven methodology to maximize the revenue of the whole system based on the data from limited samples.

- In order to preserve the privacy, the first-tier carrier/macrocell BS deploys centralized differential privacy technique before sending them to the second-tier carrier. With the centralized differential privacy protocol, the second-tier carrier can learn the statistic result of each user's transmission data amount without compromising the privacy of

the each sample's data.

- When the second-tier carrier receives the data transmission profile of users, it will decide where and how many small cell BSs to place. However, the second-tier carrier cannot know the real transmission data of users, but can only build a reference distribution that close enough to the real distribution of the transmission data requirement. In other word, for the second-tier carrier, the users' transmission data requirement is ambiguous. For instance, the second-tier carrier cannot know the precise transmission data of users per hour in the future. On the other hand, it can forecast the users' transmission data distribution in a particular hour (e.g., 12:00 am to 1:00 pm) from the historical data (historical data from 12:00 am to 1:00 pm in a month). In order to provide an optimal trade-off between service quality, availability and cost, the second-tier carrier employs data-driven approach on the collected historical data. The second-tier carrier can construct reference distribution $\mathbb{P}_0$ of users' transmission data, and ensure that the distance between the ambiguous distribution $\mathbb{P}$ and the reference distribution $\mathbb{P}_0$ is close enough.

- Based on the modeling above, we formulate the revenue maximization problem into a risk-averse two-stage stochastic problem (RA-SP). To solve this problem, we utilize ζ-structure probability metric to guarantee the distance robustness. Our proposed model solves the problem directly from the historical data without assuming/generating the true distribution of users' transmission data amount. The effectiveness of the proposed scheme is also verified by the performance evaluation results.

The rest of this work is organized as follows. In Sec. 7.2, we introduce the network model and the related model in the system. We also introduce the preliminary of differential privacy and ζ-probability metrics, and formulate the optimization problem to preserve users' transmission data privacy and maximize the NGN revenue simultaneously. In Sec. 7.3, we present how we solve the proposed formulation with optimal cut and feasible cut. Simulation

results and discussions are presented in Sec. 7.4, and the conclusion remarks are drawn in Sec. 7.5.

7.2 System Description

7.2.1 Network Configuration

Our proposed BSs deployment in the assigned region consists of the first-tier carrier with one macrocell BS and the second-tier carrier with $\mathcal{N} = \{1, 2, \cdots, i, \cdots, N\}$ small cell BSs, as shown in Fig. 7.1. As introduced in Sec. 7.1, the first-tier carrier and second-tier carrier collaborate all BSs together to satisfy all users transmission data in the fixed region. Both of the carriers have their own databases that can store the transmission data information. The database server of first-tier carrier is entitled to collect the $\mathcal{M} = \{1, 2, \cdots, j, \cdots, M\}$ users with unequal $D = \{d_1, d_2, \cdots, d_j, \cdots, d_M\}$ transmission data.

We assume the capacity of macrocell BS is C_m. In the 5G, one macrocell BS may not satisfy all users' demand due to the proliferation of the smart phones and the enormous amount of different wireless services, such as video game/face time and live. Therefore, the first-tier carrier will authorize second-tier carrier to build the small cell BSs to support communication transmission of the region. The second-tier carrier can access all users' transmission information from the database of first-tier carrier, and restore it to its own database server. After that, the server of second-tier carrier will process the data, and decide how many and which kind of smallcell BSs will be built in the assigned area.

In our NGN assumption, the adversaries could be the dishonest small cell BSs or eavesdropping attackers, who are always monitoring the information exchange between the two carrier database servers. Without enforcing any privacy preserving schemes, the adversaries can easily learn the users' transmission data profiles. The user's transmission information profile is similar to the energy profile in the smart grid. The adversaries can easily learn the

pattern of a particular user, or the scale of a business with the prior knowledge of location. For instance, the adversary will eavesdrop the resident activity pattern and speculate when the user leaves/comes back home. The adversary can also estimate how many customers are present in a hotel when it obtains the transmission data profile under the hotel's location.

To preserve the users transmission information privacy and maximum revenue for deployment of small cell BSs, it takes four steps for the first-tier carrier to deliver users' data to second-tier carrier. Firstly, the second-tier carrier sends queries about users transmission data to the first-tier carrier. Secondly, the first-tier database server collects the data information, and constructs the reference distribution $\mathbb{P}_0$. Third, the first-tier database adds noises drawn from Laplace distribution to $\mathbb{P}_0$, and establishes a noisy distribution $\mathbb{P}'_0$, which achieves ϵ-DP. Meanwhile, the first-tier database server sends the noisy distribution to the second-tier database. Fourth, the second-tier database receives the noisy data, and decides how many small cell BS to build in this region. The second-tier database knows the received data is processed by differential privacy, and builds a confident distribution set $\mathcal{D}$ which includes all the distribution that is close enough to the noisy distribution $\mathbb{P}'_0$ under certain stochastic metric, which we will explain in Sec. 7.2.2. The second-tier data base ensures that the real transmission data distribution $\mathbb{P}$ is in the confident distribution set $\mathcal{D}$. Finally, based on $\mathcal{D}$, the second-tier database decides how many and which kind small cell BSs to build to maximize its revenue.

7.2.2 Revenue Maximization Problem Formulation

First, the first-tier carrier will send the noisy information to second-tier carrier under Laplace mechanism as described in2.1.1. In our scenario, the first-tier carrier have limited historical data (e.g., 300 hours) for each users' per hour transmission data/throughput. Therefore, the d_j in $D = \{d_1, d_2, \cdots, d_j, \cdots, d_M\}$ indicates the dataset of users j's throughput per hour (with 300 records). The second-tier carrier will ask first tier the question f = "how many records of user j's throughput is among 40Mbps to 60Mbps", the first-tier

carrier will send the noisy answer to the second-tier carrier with Laplace mechanism. Similar to this, the second-tier carrier will ask another question "how many records of user j's throughput is among 60Mbps to 80Mbps",...,. Then the second-tier carrier will gather the answer of all the questions and build the noisy reference distribution of user j's throughput per hour.

The revenue maximization problem for system network can be formulated as

$$\max_{x_{ij},y} \quad -\sum_{i=1}^{N} F_i y_i + \sum_{j=1}^{M} \mathbb{E}_{\mathbb{P}_j}\left[\sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j)\right], \tag{7.1}$$

s.t.:

$$\sum_{j=1}^{M} x_{ij}(\xi_j) \leq C_i y_i, \quad i = 1, \cdots, N, \tag{7.2}$$

$$\sum_{i=1}^{N} x_{ij}(\xi_j) = d(\xi_j), \quad j = 1, \cdots, M, \tag{7.3}$$

$$r_{i,i'} \geq R_i^{\text{IN}}, \quad i \neq i', \quad i = 1, \cdots, N, \text{ and} \tag{7.4}$$

$$x_i(\xi) \geq 0, \quad y_i \in \{0, 1\} \quad \forall i, j. \tag{7.5}$$

In the problem setting, each small cell BS i associates with a fixed opening cost F_i and a capacity C_i. In the same time, there is per unit data transaction fee T_{ij}, for each user j to BS i. The binary variable y_i indicates if small cell BS i is open. The variable x_{ij} indicates the users' transmission data information from the BSs i to users j. Let $r_{i,i'}$ denote the distance between two small cell BS i and i', and R_i^{IN} denote the interference range of i. In the formulation, (7.2) indicates the total data delivery from small cell BS i should not exceed its capacity (7.3) indicates the total amount of transmission data for user j is equal to its data demand $d_j(\xi)$, and (7.4) indicates the two deployed BSs should not interfere with each other. In our assumption, the transferred data from first-tier database to second-tier carrier is processed by differential privacy. Therefore, the distribution that

the second-tier receives is the noisy reference distribution, but not the true distribution of transmission data information. To efficiently deploy the small cell BSs, the second-tier carrier would employ data-driven approach to maximize the revenue of NGN since the reference distribution transferred from first-tier database cannot present the true distribution of the users' transmission data. At the same time, the second-tier carrier supposes to satisfy the users' traffic delivery with the consideration of the cost of BSs deployment. The second-tier carrier need to predict the transmission information of the assigned area from historical *noisy* data (i.e., the first-tier database implement Laplace mechanism to original dataset), and the accuracy of the transmission data is related to the amount of historical data and the privacy budget ϵ from (2.2). Since the second-tier carrier can only get limited number of noisy historical data to study user's transmission data information, the noisy reference distribution $\mathbb{P}_0'$ that second-tier carrier learned cannot 100% represent the real distribution of the whole users transmission data, $\mathbb{P}$. Therefore, we construct the confident set $\mathcal{D}$, and consider the worst case to let $\mathbb{P} \in \mathcal{D}$ maximize the total revenue [97, 98]. The detailed formulation is described as

$$\max_{y} \quad -\sum_{i=1}^{N} F_i y_i + \min_{p_j^k} \sum_{j=1}^{M} \sum_{k=1}^{K} p_j^k \max_{x_{ij}} \sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j^k), \tag{7.6}$$

$$\text{s.t.:} \qquad (7.4) - (7.5),$$

$$\sum_{j=1}^{M} x_{ij}(\xi_j^k) \leq C_i y_i, \quad i = 1, \cdots, N, \quad k \in K, \tag{7.7}$$

$$\sum_{i=1}^{N} x_{ij}(\xi_j^k) = d_j(\xi_j^k), \quad \forall k \in K \quad \forall j \in M, \tag{7.8}$$

$$\sum_{k=1}^{K} p_j^k = 1, \quad j = 1, \cdots, M, \text{ and} \tag{7.9}$$

$$\mathbb{P} \in \mathcal{D}. \tag{7.10}$$

In the above formulation, p_j^k represents the probability of $\mathbb{P}$ for scenario ξ_j^k to happen.

For instance, $p_j^k = 20\%$ and $d_j(\xi_j^k) = 40$ Mbps indicates there is 20% probability the throughput per hour of user j is under the k-th scenario, which is $d_j(\xi_j^k) = 40$ Mbps. However, the real distribution $\mathbb{P}$ is unknown since the second-carrier only receive noisy reference distribution $\mathbb{P}_0'$ from first-carrier. The unknown $\mathbb{P}$ satisfies the constraints (7.10) in our model . The distribution distance measurement to describe confident set $\mathcal{D}$ is proposed in [10,11]. Specifically, a confidence set $\mathcal{D}$ is constructed based on predefined distance measure $d(\mathbb{P}_0', \mathbb{P})$, where $\mathbb{P}$ indicates the real unknown distribution and $\mathbb{P}_0'$ is the noisy reference distribution from sampled first-tier carrier server after employing Laplace mechanism. The distance d_ζ between two distribution $\mathbb{P}_0'$ and $\mathbb{P}$, and confidence set $\mathcal{D}$ can be defined as :

$$\mathcal{D} = \{\mathbb{P} : d_\zeta(\mathbb{P}_0', \mathbb{P}) \leq \theta\}, \tag{7.11}$$

Let $d_\zeta(\cdot, \cdot)$ denote the distance under ζ structure probability metric. θ is the tolerance of the distance between two distributions. It is correlated to the size of the sample size, i.e., the number of days that the first-tier carrier samples. It can be easily inferred that the more number of days the first-tier carrier samples (for the per day transmission data of users), the tighter $\mathcal{D}$ would be, and $\mathbb{P}$ would be more closer to $\mathbb{P}_0'$. The converge rate under ζ-probability metrics between limited noisy reference distribution (after employing Laplace Mechanism) $\mathbb{P}_0'$ and $\mathbb{P}$ is described and shown in Sec 4.4, which is shown as follows.

- For the Uniform metric:

$$Pr\big(d_U(\mathbb{P}_0', \mathbb{P} \leq \theta)\big) \geq 1 - \exp(-\frac{\theta^2}{2}M + M\epsilon), \tag{7.12}$$

- For the Kantorovich metric:

$$Pr\big(d_K(\mathbb{P}_0', \mathbb{P} \leq \theta)\big) \geq 1 - \exp(-\frac{\theta^2}{2\varnothing^2}M + M\epsilon), \tag{7.13}$$

- For the Fortet-Mourier metric:

$$Pr\big(d_{FM}(\mathbb{P}_0', \mathbb{P}) \leq \theta\big) \geq 1 - \exp\Big(-\frac{\theta^2 M}{2\varnothing^2\Lambda^2} + M\epsilon\Big). \tag{7.14}$$

Algorithm 7.1 Bender decomposition for feasible solution

1: **Input:** Transmission data of per user i.i.d drawn from the true distribution. The confident level of $\mathcal{D}$ is set to be η
2: **Output:** Objective value of the problem (7.1)
3: Obtain the reference distribution $\mathbb{P}_0$ and tolerance θ based on the historical data
4: **if** The reference distribution and true distribution are under Kantorovich metric or Fortet-Mourier metric **then**
5: Reformulate the problem to (7.22) - (7.24)
6: Feasibility check master problem of (7.22)
7: **if** Infeasible **then**
8: Generate feasible cut for master problem
9: go to line 6
10: **end if**
11: Feasibility check the subproblem of (7.22)
12: **if** Infeasible **then**
13: Generate optimal cut for subproblem
14: go to line 6
15: **end if**
16: Stop and output solution
17: **else**
18: Reformulate the problem to (7.25) - (7.27) under Uniform metric
19: Solve the problem under bender decomposition algorithm, same as line 6 to line 16
20: Output the solution
21: **end if**

7.3 Solution to The Optimization Problem

From this section, we explore how to solve (7.6) under constrains (7.4)-(7.5),(7.7)-(7.10). We assume the sample space is $\Omega = \{\xi^1, \xi^2, \cdots, \xi^K\}$. The formulation can be simplified as :

$$\max_{y} -\sum_{i=1}^{N} F_i y_i + \min_{p_j^k} \sum_{j=1}^{M} \sum_{k}^{K} p_j^k \max_{x_{ij}} \sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j^k), \tag{7.15}$$

$$\text{s.t.} \quad (7.4) - (7.5), (7.7) - (7.10),$$

$$\sum_{k=1}^{K} p_j^k = 1 \quad j = 1, \cdots, M, \text{ and} \tag{7.16}$$

$$\max_{h_i} \sum_{k=1}^{K} h_k p_j^{k0'} - \sum_{k=1}^{K} h_k p_j^k \leq \theta, \forall h_k : ||h||_\zeta \leq 1, \tag{7.17}$$

where $|h||_\zeta$ is defined according to different metrics, $p_j^{k0'}$ represents the probability of $\mathbb{P}_0'$ for scenario ξ_j^k to happen, θ is related to different distribution metric from (7.12)-(7.14). For the Kantorovich metric, $|h_i - h_j| \leq \rho(\xi^i, \xi^j)$. For the Fortet-Mourier metric, $|h_i - h_j| \leq \rho(i,j) \max\{1, \rho(\xi^i, a)^{p-1}, \rho(\xi^j, a)^{p-1}\}$. The constraints (7.16)-(7.17) can be summarized as $\sum_{k=1}^{K} a_{kv} h_i \leq b_v, v = 1, \cdots, V$, a_{kv} and b_v are different parameters under different metrics.

To reformulate the constraints , we consider the following problem:

$$\max_{h_i} \quad \sum_{k=1}^{K} h_k p_j^{k0'} - \sum_{k=1}^{K} h_k p_j^k \text{ and} \tag{7.18}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} a_{kv} h_k \leq b_v, v = 1, \cdots, V. \tag{7.19}$$

The dual problem can be formulated as:

$$\min_{u} \quad \sum_{v=1}^{V} b_v u_v \text{ and} \tag{7.20}$$

$$\text{s.t.} \quad \sum_{v=1}^{V} a_{kv} u_v \geq p_j^{k0'} - p_j^k, \forall k, \tag{7.21}$$

where u is the dual variable. Accordingly, the problem can be reformulated as follows under Kantorovich metric and Fortet-Mourier metric:

$$\max_{y} -\sum_{i=1}^{N} F_i y_i + \min_{p_j^k} \sum_{j=1}^{M} \sum_{k}^{K} p_j^k \max_{x_{ij}} \sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j^k), \tag{7.22}$$

$$\text{s.t.} \quad (7.4) - (7.5), (7.7) - (7.10),$$

$$\sum_{k=1}^{K} p_j^k = 1, \sum_{jv=1}^{V} b_{jv} u_{jv} \leq \theta_j, \forall j, \text{ and} \tag{7.23}$$

$$\sum_{jv=1}^{V} a_{kv} u_v \geq p_j^{k0'} - p_j^k, \forall k, \forall j. \tag{7.24}$$

For the Uniform metric, we can have the reformulation from the Uniform metric definition:

$$\max_{y} -\sum_{i=1}^{N} F_i y_i + \min_{p_j^k} \sum_{j=1}^{M} \sum_{k}^{K} p_j^k \max_{x_{ij}} \sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j^k), \tag{7.25}$$

$$\text{s.t.} \quad (7.4) - (7.5), (7.7) - (7.10),$$

$$\sum_{k=1}^{K} p_j^k = 1, j = 1, \cdots, M, \text{ and} \tag{7.26}$$

$$\left| (p_j^{k0'} - p_j^k) \right| \leq \theta, \forall j. \tag{7.27}$$

After reformulating the problem, we can solve the formulation (7.22) - (7.24) and (7.25) - (7.27) through Benders' decomposition algorithm. The detailed algorithm is shown in Algorithm 7.1.

The Benders' decomposition algorithm [73] is exploited to solve the problem into global optimality. Since for each scenario ξ_j^k, the second-stage optimization problem $\max_x \sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j^k)$ of (7.6) is independent of ξ_j^k for $i \neq i'$, therefore, the function (7.6) is equivalent to

$$\max_{y} -\sum_{i=1}^{N} F_i y_i + \min_{p_j^k} \max_{x_{ij}} \sum_{j=1}^{M} \sum_{k}^{K} p_j^k \sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j^k), \tag{7.28}$$

$$\text{s.t.:} \qquad (7.4) - (7.5), (7.7) - (7.10), \text{ and } (7.16) - (7.17).$$

We can calculate the second-stage minimization problem by solving its dual. The dual subproblem and dual variables λ, v associated with constraints are given by

$$\min_{\lambda,v} \sum_{k=1}^{K} \sum_{j=1}^{M} \left[\lambda_k d(\xi_j^k) - \sum_{i}^{M} c_i y_i \upsilon_k^i \right], \tag{7.29}$$

$$\text{s.t.:} \qquad \lambda_k - \upsilon_k^i \leq p_j^k T_i, \forall i, j, k, \text{ and} \tag{7.30}$$

$$\upsilon_k^i \geq 0, \forall i, k. \tag{7.31}$$

The dual variables corresponding to scenario k for constraints (7.2)-(7.4) are υ_k^i and λ_k, respectively. Because of the duality property, the optimal objective of the primal problem is equivalent to the dual problem. Obviously, the maximization operation in the primal formulation can be combined with the dual second-stage problem.

The second-stage min-max problem can be obtained as

$$\begin{aligned} \psi(y) &= \min_{p_j^k} \max_{x_{ij}} \sum_{k}^{K} \sum_{j=1}^{M} p_j^k \sum_{i=1}^{N} T_{ij} x_{ij}(\xi_j^k) \\ &= \min_{p_j^k, \lambda, v} \sum_{k=1}^{K} \sum_{j=1}^{M} \left[\lambda_k d(\xi_j^k) - \sum_{i}^{M} c_i y_i \upsilon_k^i \right], \end{aligned} \tag{7.32}$$

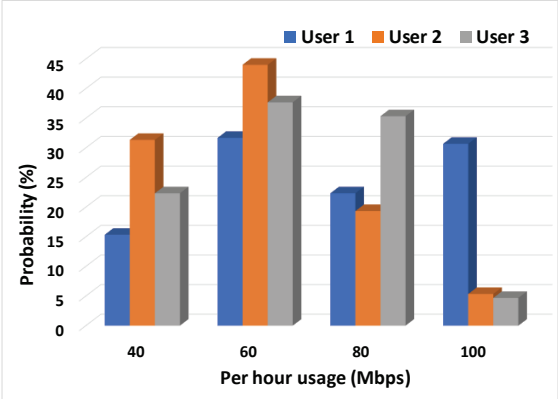

Figure 7.2: Users' per hour transmission data distribution

$$\text{s.t.:} \quad \lambda_k - \upsilon_k^i \leq p_j^k T_i, \forall i, k, j \tag{7.33}$$

$$\upsilon_k^i \geq 0, \forall i, k, \tag{7.34}$$

$$\sum_{k=1}^{K} p_j^k = 1, \forall j, \text{ and} \tag{7.35}$$

$$\mathbb{P} \in \mathcal{D}. \tag{7.36}$$

For instance, under the Uniform metric case the constrain (7.36) represents as

$$|p_j^k - p_j^{k0'}| \leq \theta_j, \forall k, \forall j, \tag{7.37}$$

where $\theta = \sqrt{2log(e^{\epsilon M}/(1-\eta))/M}$ according to (7.12).

Therefore, the (7.28) can be represented as a max-min problem with $\psi(y)$ and can be solved by applying feasibility cut and optimality cut iteratively. We denote α as the second-stage worst case, then we can solve the master problem which is reformulated as

$$\max_{y \in \{0,1\}} \sum_{i}^{N} F_i y_i + \alpha,$$

$$\text{s.t.: Feasibility cuts, and}$$

$$\text{Optimality cuts,}$$

where the feasibility cuts can be generated by L-shaped method, and the optimality cut can be generated by $\psi(y)$ accordingly.

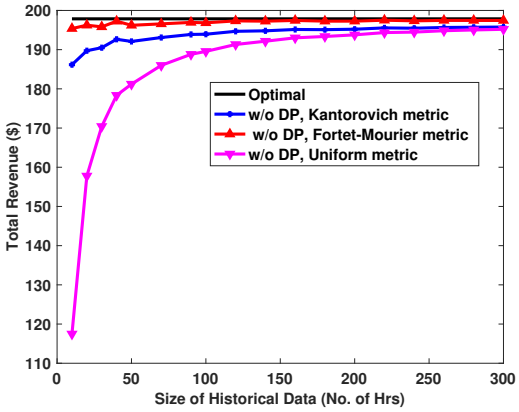

Figure 7.3: Total revenue under different metrics without different privacy.

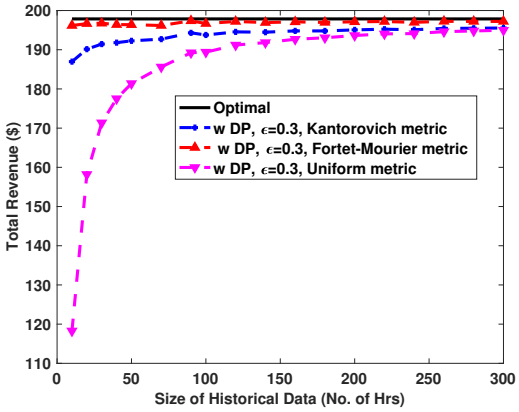

Figure 7.4: Total revenue under different metrics with privacy budget ϵ=0.3.

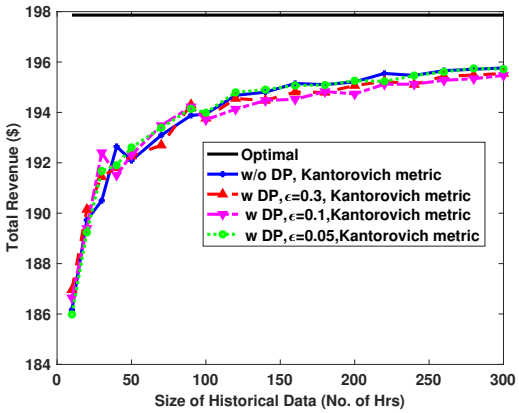

Figure 7.5: Total revenue under Kantorovich metric.

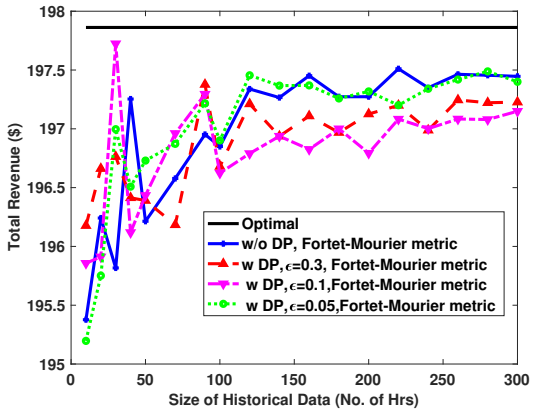

Figure 7.6: Total revenue under Fortet-Mourier metric.

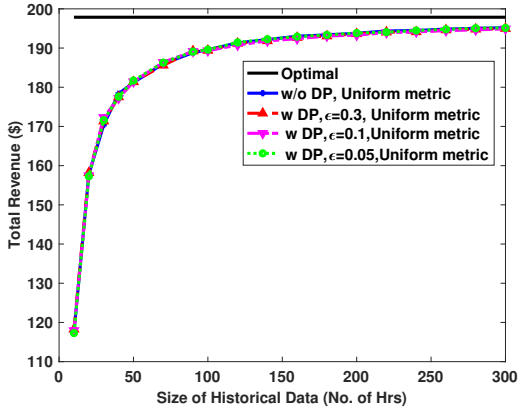

Figure 7.7: Total revenue under Uniform metric without different privacy.

7.4 Performance Evaluation

7.4.1 Simulation Setup

For illustrative purposes, we study one particular location of the city in the dataset. We choose a small region in the dataset with 3 users and 2 small cell BSs data bases. We assume the capacity of the small cell BS' data base are 150 Mbps and 180 Mbps accordingly[1]. We assume there are four scenarios of "per hour transmission data": 40 Mbps, 60 Mbps, 80 Mbps and 100 Mbps, and the system earn \$1 with delivering 1 unit data. The evaluation result is accomplished by MATLAB and GUROBI in a mac book pro with Intel Core i7 CPU of 2.7 GHz.

7.4.2 Privacy and Performance Analysis

First, we obtain 3 different distribution for the three users under 300 numbers of historical data (300 hours), which is shown in Fig 7.2. We set the confidence level as 99%, and the number of sampled hours varies from 20 to 300. First, we study the data-driven algorithm without differential privacy. The results are shown in Fig. 7.3. After collecting the data from users, the macro BS does not implement Laplace mechanism, but transfer the real users' transmission per hour to the micro BS. It can be observed that when the number of sampled hour increases, the total revenue of the system increases under all metrics. The reason is that as we have more number of sampled data, we can predict the user's transmission data distribution more precisely. When M is larger, the tolerance between two distribution θ is smaller, which stands for the confident set is more tighter around the real distribution. Also, we can observe that the gap between the simulation results under the Fortet-Mourier metric and the Kantorovich metric is very small and closer

[1]Our dataset at hand is collected hourly from 3072 4G BSs (i.e., eNBs) deployed in several southern cities in China, from September 7 to September 30, 2015. This operating data is recorded by BS interfaces, and is then delivered to the remote cloud centers for further processing or data backup. The collected data includes various operating records of 4G BSs, for example, CPU and memory usage and physical resource block (PRB) usage, which mainly belong to the network domain data in our proposed architecture. The raw data files differ in formats (e.g., file format and time stamp format).

to the optimal result, compared to Uniform metric. It indicates that the distance under Fortet-Mourier metric and Kantorovich is more tighter than Uniform metric. Besides, we study the performance with differential privacy implement in Fig 7.4. Compared to the results from 7.3, the results have same tendency. When the number of sampled hours is large enough, the total revenue with noisy data is very close to optimal revenue. It indicates that the differential privacy protocols could keep the main characteristics of the users' usage data, while preserving individual sampled user's usage privacy.

Second, we study the system revenue under different privacy budget ϵ. In Fig. 7.5 - Fig. 7.7, we set ϵ in 0.05, 0.1 and 0.3. It can be observed that the total revenue with differential privacy protocol is close to the revenue without adding noise, and lower than the optimal total revenue. However, we cannot find the relation between ϵ and total revenue is not conspicuous. The reason is that in our scenario, there are three users, i.e., three data set need to add three i.i.d from Laplace noise simultaneously. The noise may counteract with each other.

7.5 Conclusion

In this work, we study multi-tier architecture, in where the second-tier carrier deploys small cell BS to help first-carrier traffic delivery. In this architecture, we propose a novel scheme, which jointly consider user's transmission data privacy and the uncertainty of sampled users' data. We employ centralized differential privacy to keep the main characteristic of the whole dataset while preserve each individual user's information. We also propose data-driven approach to characterize the uncertainty of the users' transmission data since the limitation of the historical data set. Our approach employs ζ-structure to build a confident set between the real distribution and the reference distribution from historical data. Based on the contribution above, we formulate a risk-averse stochastic program to maximize system revenue. Through simulation, we show the system total revenue under our scheme is close to the optimal total revenue.

Chapter 8

Future Works

In the future, we will focus on big-data analysis and privacy preservation in Internet of things (IOT). IOT is the devices with which can communicate and interact with other entities over the internet and be remotely monitored and controlled. These devices are embedded with electronics, sensors and internet connectivity. To achieve this goal, the Internet of Things (IoT) considers almost any physical or logical entity given a unique ID and the capability to connect to the Internet. Meanwhile, the numerous data collected by these sensors enabled by the IoT framework is turned into knowledge and understanding of human activities. For example, human mobility traces collected through vehicles help to improve the efficiency of our transportation system and will be crucial in maximizing the potential of upcoming auto-drive vehicles. The fully connected, smart, autonomous world promises many good things for our future such as improved efficiency in using resources and energy, increased comfort level meeting personal needs, and safer and more prosperous communities.

While we are celebrating the rapid growth of IoT that has allowed us to better understand human behaviors and our surrounding, it is becoming increasingly urgent and important to understand the boundary of such techniques that, if not careful, may severely undermine our privacy. Especially for the users, while enjoying the convenience of daily life by IoT devices everywhere, they also have serious privacy concerns. For example, numerous personal data can be inferred from users' mobile phone locations and mobility traces. With smart sensing and powerful learning, we gradually feel and fear that everything is becoming overly transparent. Therefore, it is necessary to develop a smart IoT assisted network for users, and build a privacy framework for smart sensing and learning in the future.

The goal of future work is to develop a smart IoT assisted network for users, and build a privacy framework for smart sensing and learning. We would like to continue to enjoy

the benefits these technology brings to tourists. But we would also like to fully understand exactly what else can be learned from the data and make sure user specified sensitive data does not get learned. If the extra learning is not desirable, we develop new theory and algorithmic tools to maximally eliminate such concerns. This problem would involves more than just a theoretical model, as the system typically involves multiple parties/ownership, hardware components and networking elements. The proposed future work considers privacy in all elements including data gathering, aggregation, and learning, and formulate questions for upper/lower bounds of what can be achieved regarding data utility and privacy. We use two concrete examples to elaborate how this framework is adopted in the most popular IoT domains: location privacy and crowd sourcing privacy. Theoretically, we will investigate the differential privacy to the IoT scenarios as follows:

- *Smart IoT Assisted Network and Tourist Data Simulation*: The smart IoT APP for secure wireless accessing through the users' hand-held devices (e.g., smartphones, pads, wearable devices, or tablets) [99, 100] may have the user's sensitive information, and the corresponding authentication mechanisms and privacy filters. For a small users' group such as family and friends where the privacy can be partly disclosed inside, we will study how to collect and manage user's behavior history data considering different privacy preserving levels. We also plan to incorporate the proposed privacy preserving learning framework into the system, and conduct simulations over smart phones in practice to validate the designs.

- *Accuracy-Privacy Trade off of Mobile Crowd Sensing:* Privacy preservation and accuracy maximization is a critical concern in mobile crowdsensing. We will study the strategy which can be used by users to send their data under one generalized identity, increase the privacy protection, and share the resulting payoffs among cooperative users based on their individual sensing contribution.

- *Data-Driven Mobile Crowdsourcing with Users' Differential Privacy Preservation:*

Data-driven mobile crowdsourcing scheme is a promising way to help managers/operators of small business accurately estimate users' purchasing power while keeping individual user's purchasing capability information differential private, and provide better services to tourists and make more profit for thousands of retail stores. We will investigate data-driven crowdsourcing with different number of user samples and probability metrics, users' privacy preservation with different local/distributed differential privacy models, and integration of data-driven optimization and differential privacy to maximize the profit of small business.

- *Transportation Network Company (TNC) Vehicle Scheduling with Users' Location Privacy Preservation:* To better serve users in terms of transportation experiences, data-driven/learning based approach is a promising method to schedule TNC vehicles while preserving passengers' location privacy. Briefly, we will study geo-indistinguishability scheme based on differential privacy technologies. Then, TNC vehicle scheduling models will be proposed under deep reinforcement learning and data-driven method based on the passengers' obfuscated location data, respectively.

References

[1] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[2] A. Friedman, I. Sharfman, D. Keren, and A. Schuster, "Privacy-preserving distributed stream monitoring." in *NDSS*, San Diego, CA, Feb 2014.

[3] X. Jin and Y. Zhang, "Privacy-preserving crowdsourced spectrum sensing," in *Proceeding of the IEEE International Conference on Computer Communications (INFOCOM)*, 2016, pp. 1–9.

[4] G. Cormode, S. Jha, and T. Kulkarni, "Privacy at scale: local differential privacy in practice," in *2018 ACM SIGMOD/PODS International Conference on Management of Data*, Houston, TX, June 2018.

[5] E. Shi, H. Chan, E. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Annual Network & Distributed System Security Symposium (NDSS)*, San Diego, CA, February 2011.

[6] C. Dwork, "Differential privacy: a survey of results," in *International Conference on Theory and Applications of Models of Computation*, Xi'an, China, April 2008.

[7] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[8] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, Berkeley, CA, October 2013.

[9] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proceedings of the 26th USENIX Security Symposium*, Vancouver, BC, Canada, August 2017.

[10] G. C. Calafiore, "Ambiguous risk measures and optimal robust portfolios," *SIAM Journal on Optimization*, vol. 18, no. 3, pp. 853–877, October 2007.

[11] D. Klabjan, D. Simchi-Levi, and M. Song, "Robust stochastic lot-sizing by means of histograms," *Production and Operations Management*, vol. 22, no. 3, pp. 691–710, February 2013.

[12] U. SCHMOCK, "Large deviations techniques and applications," *Journal of the American Statistical Association*, no. 452, pp. 1380–1380, 2000.

[13] F. Bolley and C. Villani, "Weighted csiszar-kullback-pinsker inequalities and applications to transportation inequalities," *Annales de la Facult ÃÅe des Sciences de Toulouse*, vol. 14, pp. 331–352, 2005.

[14] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*. Methuen London, 1964.

[15] C. Zhao and Y. Guan, "Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics," *Available on Optimization Online*, 2015.

[16] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, X. Cheng, and B. Jiao, "Efficiency resource allocation for device-to-device underlay communication systems: A reverse iterative combinatorial auction based approach," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 348–358, Sep 2013.

[17] S. Bregni and L. Jmoda, "Accurate estimation of the hurst parameter of long-range dependent traffic using modified allan and hadamard variances," *IEEE Transaction on Communications*, vol. 55, no. 11, pp. 2224–2224, Nov 2007.

[18] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4569–4581, 2013.

[19] A. Rabbachin, T. Q. Quek, H. Shin, and M. Z. Win, "Cognitive network interference," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 2, pp. 480–493, Feb 2011.

[20] Z. Wang and W. Zhang, *Opportunistic spectrum sharing in cognitive radio networks*. Springer, 2015.

[21] F. Wang, M. Krunz, and S. Cui, "Price-based spectrum management in cognitive radio networks," *IEEE Journal of selected topics in signal processing*, vol. 2, no. 1, pp. 74–87, Feb 2008.

[22] Y.-C. Liang, K.-C. Chen, G. Y. Li, and P. Mahonen, "Cognitive radio networking and communications: An overview," *IEEE transactions on vehicular technology*, vol. 60, no. 7, pp. 3386–3407, Sep 2011.

[23] S. EZarrin and T. J. Lim, "Throughput-sensing tradeoff of cognitive radio networks based on quickest sensing," in *Proceeding of the IEEE International Conference on Communications (ICC'11)*, Kyoto, Japan, June 2011.

[24] Y. Zeng, Y.-C. Liang, A. T. Hoang, and R. Zhang, "A review on spectrum sensing for cognitive radio: challenges and solutions," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, p. 381465, Dec 2010.

[25] M. Clark and K. Psounis, "Can the privacy of primary networks in shared spectrum be protected?" in *Proceeding of the IEEE International Conference on Computer Communications (INFOCOM)*, San Fransisco, CA, April 2016.

[26] A. Robertson, J. Molnar, and J. Boksiner, "Spectrum database poisoning for operational security in policy-based spectrum operations," in *IEEE Military Communications Conference*, San Diego, CA, November 2013.

[27] B. Bahrak, S. Bhattarai, A. Ullah, J.-M. Park, J. Reed, and D. Gurney, "Protecting the primary users operational privacy in spectrum sharing," in *IEEE International Symposium on Dynamic Spectrum Access Networks*, Mclean, VA, April 2014.

[28] Z. Gao, H. Zhu, S. Li, S. Du, and X. Li, "Security and privacy of collaborative spectrum sensing in cognitive radio networks," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 106–112, December 2012.

[29] J. Liu, C. Zhang, H. Ding, H. Yue, and Y. Fang, "Policy-based privacy-preserving scheme for primary users in database-driven cognitive radio networks," in *Proceeding of the IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, December 2016.

[30] X. Fu, B. Graham, R. Bettati, and W. Zhao, "On effectiveness of link padding for statistical traffic analysis attacks," in *23rd International Conference on Distributed Computing Systems*, May 2003.

[31] T. Bonald, L. Massoulié, A. Proutiere, and J. Virtamo, "A queueing analysis of max-min fairness, proportional fairness and balanced fairness," *Queueing systems*, vol. 53, no. 1, pp. 65–84, 2006.

[32] Y. Chen and H.-S. Oh, "A survey of measurement-based spectrum occupancy modeling for cognitive radios," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 848–859, Jan 2016.

[33] Y. Saleem and M. H. Rehmani, "Primary radio user activity models for cognitive radio networks: A survey," *Journal of Network and Computer Applications*, vol. 43, pp. 1–16, Aug 2014.

[34] X. Xing, T. Jing, W. Cheng, Y. Huo, and X. Cheng, "Spectrum prediction in cognitive radio networks," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 90–96, Apr 2013.

[35] M. Höyhtyä, A. Mämmelä, M. Eskola, M. Matinmikko, J. Kalliovaara, J. Ojaniemi, J. Suutala, R. Ekman, R. Bacchus, and D. Roberson, "Spectrum occupancy measurements: A survey and use of interference maps," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2386–2414, January 2016.

[36] J. Lundén, S. A. Kassam, and V. Koivunen, "Robust nonparametric cyclic correlation-based spectrum sensing for cognitive radio," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 38–52, January 2010.

[37] S. Gong, P. Wang, W. Liu, and W. Zhuang, "Performance bounds of energy detection with signal uncertainty in cognitive radio networks," in *INFOCOM, Proceedings IEEE*, 2013, pp. 2238–2246.

[38] L. Zhang, M. Xiao, G. Wu, S. Li, and Y.-C. Liang, "Energy-efficient cognitive transmission with imperfect spectrum sensing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1320–1335, May 2016.

[39] F. Akhtar, M. H. Rehmani, and M. Reisslein, "White space: Definitional perspectives and their role in exploiting spectrum opportunities," *Telecommunications Policy*, vol. 40, no. 4, pp. 319–331, 2016.

[40] M. Monemi, M. Rasti, and E. Hossain, "On characterization of feasible interference regions in cognitive radio networks," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 511–524, 2016.

[41] Y. T. Hou, Y. Shi, and H. D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 146–155, January 2008.

[42] J. Tang, S. Misra, and G. Xue, "Joint spectrum allocation and scheduling for fair spectrum sharing in cognitive radio wireless networks," *Computer Networks (Elsevier) Journal*, vol. 52, no. 11, pp. 2148–2158, August 2008.

[43] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu, "Impact of interference on multi-hop wireless network performance," in *Proc. of Mobile Computing and Networking (Mobicom'03)*, San Diego, CA, September 2003.

[44] H. Li, Y. Cheng, C. Zhou, and P. Wan, "Multi-dimensional conflict graph based computing for optimal capacity in MR-MC wireless networks," in *Proc. of International Conference on Distributed Computing Systems (ICDCS)*, Genoa, Italy, June 2010.

[45] A. N. Kadhim, F. Hajiaghajani, and M. Rasti, "On selecting duplex-mode and resource allocation strategy in full duplex d2d communication," Tehran, Iran, May 2017.

[46] M. Amjad, F. Akhtar, M. H. Rehmani, M. Reisslein, and T. Umer, "Full-duplex communication in cognitive radio networks: A survey," *IEEE Communications Surveys & Tutorials*, 2017.

[47] I. Akyildiz, W. Lee, M. Vuran, and M. Shantidev, "Next generation/ dynamic spectrum access/ cognitive radio wireless networks: a survey," *Computer Networks (Elsevier) Journal*, vol. 50, no. 4, pp. 2127–2159, September 2006.

[48] IEEE 802.22-2011(TM) Standard for Cognitive Wireless Regional Area Networks (RAN) for Operation in TV Bands, July 2011.

[49] FCC, "Spectrum policy task force report," Report of Federal Communications Commission, Et docket No. 02-135, November 2002.

[50] M. Pan, P. Li, Y. Song, Y. Fang, P. Lin, and S. Glisic, "When spectrum meets clouds: Optimal session based spectrum trading under spectrum uncertainty," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 3, pp. 615–627, March 2014.

[51] L. Duan, J. Huang, and B. Shou, "Cognitive mobile virtual network operator: Investment and pricing with supply uncertainty," in *Proc. of IEEE Conference on Computer Communications, INFOCOM 2010*, San Diego, CA, March 2010.

[52] X. Li, H. Ding, M. Pan, Y. Sun, and Y. Fang, "Users first: Service-oriented spectrum auction with a two-tier framework support," *IEEE Journal on Selected Areas in Communications: Spectrum Sharing and Aggregation for Future Wireless Networks*, vol. 34, no. 11, pp. 2999–3013, November 2016.

[53] Q. Huang, Y. Tao, and F. Wu, "Spring: A strategy-proof and privacy preserving spectrum auction mechanism," in *Proceeding of IEEE International Conference on Computer Communications (INFOCOM)*, Turian, Italy, April 2013.

[54] M. Li, P. Li, L. Guo, and X. Huang, "PPER: Privacy-preserving economic-robust spectrum auction in wireless networks," in *IEEE Conference on Computer Communications (INFOCOM)*, Kowloon, April 2015, pp. 909–917.

[55] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[56] R. Zhu, Z. Li, F. Wu, K. Shin, and G. Chen, "Differentially private spectrum auction with approximate revenue maximization," in *Proceedings of ACM international symposium on mobile ad hoc networking and computing, ACM MobiHoc*, 2014, pp. 185–194.

[57] C. zhao and Y. Yuan, "Data-driven stochastic unit commitment for integrating wind generation," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2587–2596, 2016.

[58] S. M. Errapotu, J. Wang, Z. Lu, W. Li, M. Pan, and Z. Han, "Bidding privacy preservation for dynamic matching based spectrum trading," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM'16)*, Washington, D.C., USA, December 2016.

[59] J. Sun, R. Zhang, J. Zhang, and Y. Zhang, "Pristream: Privacy-preserving distributed stream monitoring of thresholded percentile statistics," in *Proceeding of the IEEE International Conference on Computer Communications (INFOCOM)*, 2016, pp. 1–9.

[60] S. Gong, P. Wang, and W. Liu, "Spectrum sensing under distribution uncertainty in cognitive radio networks," in *IEEE International Conference on Communications (ICC)*, Ottawa, ON, USA, June 2012.

[61] IEEE Spectrum, 2018. [Online]. Available: https://spectrum.ieee.org/energywise/energy/the-smarter-grid/17-billion-modernization-plan-for-puerto-rico-is-released

[62] Next Grid: Illinois Utility of the Future Study, 2018. [Online]. Available: https://nextgrid.illinois.gov/about.html

[63] X. Lou, R. Tan, D. K. Yau, and P. Cheng, "Cost of differential privacy in demand reporting for smart grid economic dispatch," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, May 2017.

[64] J. Kamto, L. Qian, J. Fuller, J.attia, and Y. Qian, "Key distribution and management for power aggregation and accountability in advance metering infrastructure," in *IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, Tainan, Taiwan, November 2012.

[65] T. Baumeister, "Adapting pki for the smart grid," in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Brussels, Belgium, October 2011.

[66] M. M. Fouda, Z. M. Fadlullah, N. Kato, R. Lu, and X. S. Shen, "A lightweight message authentication scheme for smart grid communications," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 675–685, December 2011.

[67] A. R. Metke and R. L. Ekl, "Security technology for smart grid networks," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 99–107, June 2010.

[68] M. R. Asghar, G. Russello, B. Crispo, and M. Ion, "Supporting complex queries and access policies for multi-user encrypted databases," in *Proceedings of the ACM Cloud Computing Security Workshop, Co-located with CCS*, Berlin, Germany, November 2013.

[69] O. Vukovic, G. Dan, and R. B. Bobba, "Confidentiality-preserving obfuscation for cloud-based power system contingency analysis," in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Vancouver, BC, Canada, December 2013.

[70] Z. Yang, S. Zhong, and R. N. Wright, "Privacy-preserving classification of customer data without loss of accuracy," in *Proceedings of the 2005 SIAM International Conference on Data Mining*, Newport Beach, CA, April 2005.

[71] J. Wang, Y. Gong, L. Qian, R. Jaentti, M. Pan, and Z. Han, "Primary users' operational privacy preservation via data-driven 0ptimization," in *2017 IEEE Global Communications Conference (GLOBECOM)*, Singapore, December 2017.

[72] C. Zhao and Y. Guan, "Data-driven stochastic unit commitment for integrating wind generation," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2587–2596, July 2016.

[73] A. M. Geoffrion, "Generalized benders decomposition," *Journal of optimization theory and applications*, vol. 10, no. 4, pp. 237–260, October 1972.

[74] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, July 2012.

[75] M. Mangili, F. Martignon, S. Paris, and A. Capone, "Bandwidth and cache leasing in wireless information-centric networks: A game-theoretic study," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 679–695, January 2017.

[76] S. Puglisi, J. Parra-Arnau, J. Forné, and D. Rebollo-Monedero, "On content-based recommendation and user privacy in social-tagging systems," *Computer Standards & Interfaces*, vol. 41, pp. 17–27, September 2015.

[77] "Cisco visual networking index: Forecast and methodology, 2016 − 2021," Cisco, 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

[78] K. Wang, H. Li, F. R. Yu, and W. Wei, "Virtual resource allocation in software-defined information-centric cellular networks with device-to-device communications and imperfect CSI," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10 011–10 021, December 2016.

[79] M. Hajimirsadeghi, N. B. Mandayam, and A. Reznik, "Joint caching and pricing strategies for popular content in information centric networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 3, pp. 654–667, March 2017.

[80] L. A. Adamic and B. A. Huberman, "Zipf's law and the internet." *Glottometrics*, vol. 3, no. 1, pp. 143–150, 2002.

[81] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forne, and D. Rebollo-Monedero, "Privacy-preserving enhanced collaborative tagging," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 180–193, January 2014.

[82] Z. Zhou, M. Dong, K. Ota, and Z. Chang, "Energy-efficient context-aware matching for resource allocation in ultra-dense small cells," *IEEE Access*, vol. 3, no. 9, pp. 1849–1860, 2015.

[83] J. Zheng, Y. Wu, N. Zhang, H. Zhou, Y. Cai, and X. Shen, "Optimal power control in ultra-dense small cell networks: A game-theoretic approach," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4139–4150, 2017.

[84] T. S. Rappaport, W. Roh, and K. Cheun, "Wireless engineers long considered high frequencies worthless for cellular systems. they couldn't be more wrong," *IEEE Spectrum*, vol. 51, no. 9, pp. 34–58, 2014.

[85] A. Ericsson, "5g radio access-research and vision," *Ericsson White Paper 284 23-3204 Uen*, 2013.

[86] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[87] B. Galkin, J. Kibilda, and L. A. DaSilva, "Deployment of uav-mounted access points according to spatial user locations in two-tier cellular networks," in *Wireless Days (WD)*, Toulouse, France, March 2016.

[88] M. Deruyck, W. Joseph, and L. Martens, "Power consumption model for macrocell and microcell base stations," *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 3, pp. 320–333, 2014.

[89] C. S. Chen, V. M. Nguyen, and L. Thomas, "On small cell network deployment: A comparative study of random and grid topologies," in *IEEE Vehicular Technology Conference (VTC Fall)*, QC, Canada, September 2012.

[90] I. Bilogrevic, M. Jadliwala, and J.-P. Hubaux, "Security issues in next generation mobile networks: Lte and femtocells," in *2nd international femtocell workshop*, No. EPFL-POSTER-149153.

[91] 3GPP TS 23.003 v4.9.0. Numbering, addressing and identification. Visited on 07.05.2018.

[92] 3GPP TS 23.003 v8.6.0. Numbering, addressing and identification. Visited on 07.05.2018.

[93] J. Mirkovic, M. Robinson, P. Reiher, and G. Oikonomou, "Distributed defense against DDOS attacks," *University of Delaware CIS Department technical report CIS-TR-2005-02*, pp. 1–12, 2005.

[94] C. Papadopoulos, R. Lindell, J. Mehringer, A. Hussain, and R. Govindan, "Cossack: coordinated suppression of simultaneous attacks," in *Proceedings DARPA Information Survivability Conference and Exposition*, Washington DC, April 2003.

[95] B. Waters, A. Juels, J. A. Halderman, and E. W. Felten, "New client puzzle outsourcing techniques for DoS resistance," in *Proceedings of the 11th ACM conference on Computer and communications security*, Washington DC, October 2004.

[96] M. Guri, Y. Mirsky, and Y. Elovici, "9-1-1 DDoS: Threat, analysis and mitigation," *arXiv preprint arXiv:1609.02353*, 2016.

[97] J. Wang, X. Zhang, H. Zhang, H. Lin, H. Tode, M. Pan, and Z. Han, "Data-driven optimization for utility providers with differential privacy of users' energy profile," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Singapore, December 2018.

[98] J. Wang, S. M. Errapotu, Y. Gong, L. Qian, R. Jäntti, M. Pan, and Z. Han, "Data-driven optimization based primary users' operational privacy preservation," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 2, pp. 357–367, October 2018.

[99] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, Oct 2016.

[100] H. Nakano, Y. Tanigawa, and H. Tode, "Dynamic adaptation to environmental changes of optical virtual networking and cloud computing systems for tightly coupling big data

and peripheral computer resources," in *IEEE Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, 2018.