PRIVACY CHARACTERIZATION AND QUANTIFICATION IN DATA PUBLISHING

By

Mohamed Hossam Afifi Ibrahim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Electrical Engineering — Doctor of Philosophy

2021

ABSTRACT

PRIVACY CHARACTERIZATION AND QUANTIFICATION IN DATA PUBLISHING

By

Mohamed Hossam Afifi Ibrahim

The increasing interest in collecting and publishing large amounts of individuals' data to public for purposes such as medical research, market analysis and economical measures has created major privacy concerns about their sensitive information. To deal with these concerns, many Privacy-Preserving Data Publishing (PPDP) schemes have been proposed in literature. However, they lack a proper privacy characterization. As a result, the existing schemes fail to provide reliable privacy loss quantification metrics and thus fail to correctly model the utility-privacy tradeoff.

In this thesis, we first present a novel multi-variable privacy characterization model. Based on this model, we are able to analyze the prior and posterior adversarial beliefs about attribute values of individuals. Then we show that privacy should not be measured based on one metric. We demonstrate how this could result in privacy misjudgment. We propose two different metrics for quantification of privacy loss. Using these metrics and the proposed framework, we evaluate some of the most well-known PPDP techniques.

The proposed metrics and data publishing framework are then used to build a negotiation-based data disclosure model to jointly address the utility requirements of the Data User (DU) and the privacy and, possibly, the monetary requirements of the Data Owner (DO). The data utility is re-defined based on the DU's rather than the DO's perspective. Based on the proposed model, we present two data disclosure scenarios that satisfy a given privacy constraint while achieving the DU's required data utility level. The variation in a DO's flat

or variable monetary rate objective motivates the data disclosure scenarios. This model fills the gap between the existing theoretical work and the ultimate goal of practicality.

The data publisher is required to provide guarantees that users' records cannot be de-identified from datasets. This reflects directly on the levels of data generalization and techniques by which data is anonymized. While Machine Learning (ML), one of the most revolutionary technologies nowadays, relies mainly on data, it is unfortunate that the more generalized the data is, the less accurate the ML model becomes. Although this is a well understood fact, we lack a model that quantifies such degradation in ML models' accuracy, as a consequence to the privacy constraints. To model this tradeoff, we provide the first framework to quantify, not only the privacy losses in data publishing, but also the utility losses in machine learning applications as a result of meeting the privacy constraints.

To further expand our research and reflect its applicability to real industry applications, the proposed tradeoff management framework is then applied on a large-scale employee dataset from Barracuda Networks, a leader cybersecurity company. A privacy-preserving Account Takeover (ATO) detection algorithm is then proposed to predict the fraudulence of email account logins and thus detect possible ATO attacks. The results express variations in models' accuracy in binary classification of logins when trained on different datasets that satisfy different privacy constraints. The proposed framework enables a data owner to quantitatively manage the utility-privacy tradeoff and provide deeper insights about the value of the released data as well as the potential privacy losses upon publishing.

I dedicate this work to my dear parents, wife, and kids.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my advisor and mentor Dr. Jian Ren. Apart from all the invaluable knowledge that he has passed to me, I learned a lot from him as a person. I will never forget when he visited and cooked for me when I had an unexpected surgery. Such little actions show what kind of person one is. I cannot thank him enough for everything he has taught me and will always be grateful.

I would like to thank Dr. Tongtong Li, Dr. Kalyanmoy Deb, and Dr. Richard Enbody for serving on my committee. Your support and feedback were a great value-added to this dissertation. I would also like to express my appreciation to Dr. Tongtong Li for all what she taught me as a student and as a TA for her class. Also, thank you for all the lunches and dinners we had together.

I must thank my labmates, Kai and Ehab. You were both the coolest people to ever work with and I have learned so much from both of you. I will never forget these fun times.

I cannot thank my dear father enough for pushing me in the direction of pursuing a Ph.D. You have always believed in me and motivated me in every single possible way. My beloved mother, your constant prayers and comforting words on our daily phone calls are the reasons why I am where I am today. My sisters, Rana and Nouran, the true gems in my life. You've always been there for me whenever needed. How kind, loving, and caring you've always been have always inspired me to be a better person.

Nada, my wife and my angel, without your love and support I wouldn't have been able to make it. You never saved any effort to provide the suitable atmosphere for me to work. My daughters, Tamara and Lana, you might still be young, but one day I'll explain to you how much of an inspiration you were to me. I hope I can always make you proud.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

LIST OF ALGORITHMS . . . . . **xii**

Chapter 1 Introduction **1**
1.1 Overview 1
1.2 Related Work 2
1.3 Summary of Contributions 4
1.3.1 Privacy Characterization and Quantification in Data Publishing . . . 4
1.3.2 UBNB-PPDP: Utility-Boosting Negotiation-Based Privacy Preserving Data Publishing 5
1.3.3 Privacy Preserving Data Publishing for Machine Learning Applications 6
1.3.4 Privacy Preserving ATO Detection 7
1.4 Thesis Organization 8

Chapter 2 The Proposed Publishing Model and Privacy Characterization 9
2.1 Introduction 9
2.2 Preliminaries 10
2.2.1 Data Publishing 10
2.2.2 Attacks on Datasets 14
2.2.3 Analysis of the Existing PPDP Schemes 15
2.2.3.1 k-anonymity 15
2.2.3.2 ℓ-diversity 17
2.2.3.3 t-closeness 20
2.3 Generalization Model 22
2.4 The Adversarial Prior and Posterior Beliefs 24
2.4.1 The Adversarial Prior Belief 24
2.4.2 The Adversarial Posterior Belief 25
2.5 Defining Privacy Loss 26
2.6 Summary 29

Chapter 3 The proposed Privacy Quantification Metrics 30
3.1 Introduction 30
3.2 The Intuition Behind the Proposed Metrics 31
3.3 The Proposed Privacy Quantification Metrics 32
3.3.1 The Distribution Privacy Loss Metric 33
3.3.2 The Entropy Privacy Loss Metric 37
3.4 Empirical Analysis and Simulation Results 44

3.4.1 Empirical Analysis . . . 45
3.4.2 Simulation Results . . . 52
3.5 Summary . . . 55

Chapter 4 Utility-Boosting Negotiation-Based PPDD . . . 59
4.1 Introduction . . . 59
4.2 The Proposed Utility Quantification Metrics . . . 61
4.3 Utility-Privacy Tradeoff Characterization . . . 65
4.3.1 Problem Formulation . . . 66
4.3.2 The Data Generalization Management . . . 69
4.4 The Proposed UBNB-PPDD Model . . . 69
4.5 The Proposed UBNB-PPDD Protocol . . . 74
4.5.1 The Flat Rate UBNB-PPDD . . . 75
4.5.2 The Variable Rate UBNB-PPDD . . . 76
4.6 Empirical Analysis and Simulation Results . . . 77
4.6.1 Empirical Analysis . . . 77
4.6.2 Simulation Results . . . 79
4.7 Summary . . . 81

Chapter 5 Privacy Preserving Data Publishing for Machine Learning Applications . . . 85
5.1 Introduction . . . 85
5.2 Related Work . . . 88
5.3 Preliminaries . . . 90
5.4 The Privacy Loss Metrics . . . 92
5.5 Machine Learning Algorithms and their Evaluation . . . 94
5.5.1 Training the Classifier . . . 94
5.5.2 Machine Learning Algorithms . . . 94
5.5.3 Evaluation Metrics . . . 96
5.6 The Utility Loss Metric . . . 96
5.7 Privacy Preserving ATO Detection . . . 98
5.7.1 Account Takeover Attacks . . . 99
5.7.2 Data Exploration . . . 100
5.7.3 Feature Extraction . . . 102
5.7.4 Training on Original and Generalized Data . . . 103
5.7.5 Utility-Privacy Tradeoff Management . . . 105
5.7.6 Ethics . . . 110
5.8 Summary . . . 111

Chapter 6 Conclusions and Future Work . . . 112
6.1 Conclusions . . . 112
6.2 Future Work . . . 113

BIBLIOGRAPHY . . . 116

LIST OF TABLES

Table 2.1: Original table 16

Table 2.2: A 3-anonymous version 16

Table 2.3: Original dataset 19

Table 2.4: A 3-diverse version 19

Table 2.5: 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease 20

Table 3.1: Original dataset 49

Table 3.2: 4-anonymous impatient micro-data 49

Table 3.3: 3-diverse impatient micro-data 49

Table 3.4: Original dataset 50

Table 3.5: 4-anonymous, 2-diverse dataset 50

Table 3.6: Description of adults census database 53

Table 4.1: Original dataset 78

Table 4.2: 4-anonymous impatient micro-data 78

Table 5.1: Description of Microsoft azure active directory dataset 101

Table 5.2: Sample raw-data 104

Table 5.3: Sample IP-data 105

Table 5.4: Feature sets from different generalized datasets 107

Table 5.5: Model results of training on original and generalized datasets 109

Table 5.6: Tradeoff results when training on original vs. generalized datasets . . . 109

LIST OF FIGURES

Figure 1.1: The thesis milestones 7

Figure 2.1: Our privacy and utility characterization approach 24

Figure 2.2: Definition of privacy loss 27

Figure 3.1: Comparison of different statistical distance metrics 33

Figure 3.2: Example to show insufficiency of distribution loss for privacy quantification 46

Figure 3.3: Example to show the distinction of the proposed privacy metrics 46

Figure 3.4: Evaluation of a table satisfying 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $e = 2$ 56

Figure 3.5: Evaluation of a table satisfying 0.5-closeness, 7-diversity, and $k \geq 7$-anonymity at $e = 3$ 57

Figure 3.6: Loss at different sets of QIDs 58

Figure 4.1: Negotiation-based data disclosure model 60

Figure 4.2: Definition of utility loss 61

Figure 4.3: Minimum utility loss privacy-constrained data disclosure 66

Figure 4.4: Minimum privacy loss utility-constrained data disclosure 67

Figure 4.5: Utility-privacy loss constrained data disclosure 67

Figure 4.6: Privacy and utility tradeoffs 68

Figure 4.7: Utility pattern 70

Figure 4.8: Example showing the negotiation using the utility pattern $\mathcal{U}$ 71

Figure 4.9: Diagram showing utility-privacy tradeoff and the DO's profit 74

Figure 4.10: The UBNB-PPDD protocols 75

Figure 4.11: Evaluating privacy and utility losses of a table satisfying 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $e = 2$ 83

Figure 4.12: Privacy and utility losses at different sets of QIDs 84

Figure 4.13: Utility-privacy tradeoff for different attributes of interest at $e = 3, 5$ and 7 84

Figure 5.1: The proposed system model 88

Figure 5.2: Training the classifier 95

Figure 5.3: The proposed ATO detection Model 103

Figure 5.4: Feature importance of original-data trained model 106

Figure 5.5: Feature importance of subnet-generalized-data trained model 107

Figure 5.6: Feature importance of country-generalized-data trained model 108

LIST OF ALGORITHMS

Algorithm 1: Flat rate UBNB-PPDD 76

Algorithm 2: Variable rate UBNB-PPDD 77

Algorithm 3: The proposed iterative learning algorithm 98

Chapter 1

Introduction

1.1 Overview

Nowadays, datasets are considered a valuable source of information for the medical research, market analysis and economical measures. These datasets can include information about individuals that contain social, medical, statistical, and customer data. Many organizations, companies and institutions publish privacy related datasets. While the shared dataset gives useful societal information to researchers, it also creates security risks and privacy concerns to the individuals whose data are in the table. To avoid possible identification of individuals from records in published data, uniquely identifying information such as names and social security numbers are generally removed from the table. While the obvious personal identifiers are removed, the quasi-identifiers such as zip-code, age, and gender may still be used to uniquely identify a significant portion of the population since the released data makes it possible to infer or limit the available options of individuals than would be possible without releasing the table. In fact, [1, 2] showed that by correlating this data with the publicly available side information, such as information from voter registration list for Cambridge Massachusetts, medical visits about many individuals could be easily identified. This study estimated that 87% of the population of the United States could be uniquely identified using quasi-identifiers through side information based attacks, including the medical records of the

governor of Massachusetts in the medical data.

Research on data privacy has purely been focused on privacy definitions, such as k-anonymity, ℓ-diversity, and t-closeness. These models only consider minimizing the amount of privacy loss without directly measuring what the adversary may learn. There is a motivation to find consistent measurements of how much information is leaked to an adversary by publishing a dataset. Therefore, a privacy characterization model that is able to properly analyze the existing data publishing techniques is essential. Moreover, while privacy rules are inevitable, data owners will always seek a data disclosure model that can maximize data utility within the frame of the imposed privacy rules. A data disclosure model that jointly addresses the utility requirements of the Data User (DU) and the privacy and, possibly, the monetary requirements of the Data Owner (DO).

1.2 Related Work

The spate of privacy related incidents has spurred a long line of research in privacy notions for data publishing and analysis, such as k-anonymity, ℓ-diversity and t-closeness, to name a few [2–11]. A table satisfies k-anonymity if each quasi-identifier attribute in the table is indistinguishable from at least $k-1$ other quasi-identifier attributes; such a table is called a k-anonymous table. While k-anonymity protects identity disclosure of individuals by linking attacks, it is insufficient to prevent attribute disclosure with side information. By combining the released data with side information, it makes it possible to infer the possible sensitive attributes corresponding to an individual. Once the correspondence between the identifier and the sensitive attributes is revealed for an individual, it may harm the individual and the distribution of the entire table. To deal with this issue, ℓ-diversity was introduced in [4].

ℓ-diversity requires that the sensitive attributes contain at least ℓ *well-represented* values in each equivalence class. As stated in [5], ℓ-diversity has two major problems. One, is that it limits the adversarial knowledge, while it is possible to acquire knowledge of a sensitive attribute from generally available global distribution of the attribute. Another problem is that all attributes are assumed to be categorical, which assumes that the adversary either gets all the information or gets nothing for a sensitive attribute.

In [5], authors propose a privacy notion called t-closeness. They first formalize the idea of global background knowledge and propose the base model t-closeness. This model requires the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This distance was introduced to measure the information gain between the posterior belief and prior belief through the Earth Mover Distance (EMD) metric [12], which is represented as the information gain for a specific individual over the entire population. However, the value t is an abstract distance between two distributions that does not have any intuitive relation to the privacy loss. Moreover, as we show in this thesis, the distance between two distributions cannot be easily quantified by a single measurement. t-closeness also has many limitations that will be described later. The state of the art PPDP techniques will be further analyzed in more details in section 2.2.3. Our work is mainly focused on datasets where only a single attribute is considered sensitive. In [13–19], models for anonymizing datasets with multiple sensitive attributes are proposed.

Furthermore, many approaches have been proposed in literature to address the utility-privacy tradeoff. Data anonymization under privacy and utility constraints has been introduced in [20–22]. A threat model for protecting against specified inferences has been explored in Pufferfish privacy [23] and Blowfish privacy [24]. Information theoretic approaches such

as quantifying the mutual information, have been introduced in [25–30]. These approaches provide a better modeling of the tradeoff by incorporating the statistics of the dataset, assuming reasonable restrictions on the capabilities of the adversary, and modeling the side information. However, as shown in [31], since they require learning the parameters of a coding scheme by minimizing a loss function, these approaches lack the practicality when dealing with real data. A machine learning approach to address the utility-privacy tradeoff was presented in [32]. However, due to the lack of a proper utility and privacy characterization, we believe the used metrics do not provide a justifiable quantification of both utility and privacy losses where they deal with the quantification as a single dimension problem. Thus, the existing techniques of modeling the tradeoff are mostly either inapplicable or just intuitive rather than practical.

1.3 Summary of Contributions

A summary of our thesis milestones and contributions is shown in Fig. 1.1.

1.3.1 Privacy Characterization and Quantification in Data Publishing

In this thesis, we begin by introducing our novel data publishing framework. The proposed framework consists of two parts. First, we model attributes in a dataset as a multi-variable model. Based on this model, we are able to re-define the prior and posterior adversarial beliefs about attribute values of individuals. Then we characterize privacy of these individuals based on the privacy risks attached with combining different attributes. This model is indeed a more precise model to describe privacy risk of publishing datasets.

We contend that without a proper privacy characterization and quantification framework it is impossible to address the potential privacy loss issues. Our proposed framework can serve as an enabler to address the privacy losses for big data publishing [33, 34]. Having a solid framework for privacy characterization and quantification is indeed the first step on the track to tackle these issues. It also enables the big data publisher to determine the tradeoff between data utility and privacy losses.

For a given dataset, before it is released, we want to determine to what extent we can achieve privacy. Therefore, we introduce a new set of privacy quantification metrics to measure the gap between prior information belief and posterior information belief of an adversary, from both local and global perspectives. Specifically, we introduce two privacy loss metrics: *distribution loss* and *entropy loss*. We discuss the rationale for these two metrics and illustrate their advantages through examples and simulations on US Adult Census dataset. We show how considering only one metric ignoring the effect of the other leads to insufficient privacy quantification. In fact, we show that distribution and entropy loss are two distinct metrics. We show that a minimized distribution loss between sensitive attribute values distributions of the original and the published datasets does not essentially achieve the minimum entropy loss that an adversary could gain. We believe that for a published dataset to achieve better privacy, both metrics have to be taken into consideration.

1.3.2 UBNB-PPDP: Utility-Boosting Negotiation-Based Privacy Preserving Data Publishing

Based on our privacy characterization and quantification framework we propose a practical data disclosure model that introduces data utility as a function of the DU's requirements,

namely, attributes of interest. The model incorporates a negotiation process between the DO and the DU in order to reach a data disclosure deal. The DU represents their requirements as utility patterns of the attributes of interest while the DO's requirements are represented as generalization policies. Based on this model we propose two Utility-Boosting Negotiation-Based Privacy Preserving Data Disclosure (UBNB-PPDD) protocols that are guided by the DO's objective. The protocols provide a set of rules for the communication sessions between the DO and the DU in order to reach a data disclosure deal. The first protocol manages a negotiation process to disclose any generalized dataset that matches the DU's utility requirements and meanwhile satisfies the DO's privacy constraints. In the second protocol, the DO links the utility level of the disclosed dataset to a profit function.

1.3.3 Privacy Preserving Data Publishing for Machine Learning Applications

The appealing need for privacy measures has imposed various constraints on data publishing. The data publisher is required to provide guarantees that users' records cannot be de-identified from datasets. This reflects directly on the levels of data generalization and techniques by which data is anonymized. While Machine Learning (ML), the most revolutionary technology nowadays, relies mainly on data, it is unfortunate that the more generalized the data is, the less accurate the ML model becomes. Although this is a well understood fact, we lack a model that quantifies such degradation in ML models' accuracy, as a consequence to the privacy constraints. In this thesis, we provide the first framework to quantify, not only the privacy losses in data publishing, but also the utility losses in machine learning applications as a result of meeting the privacy constraints.

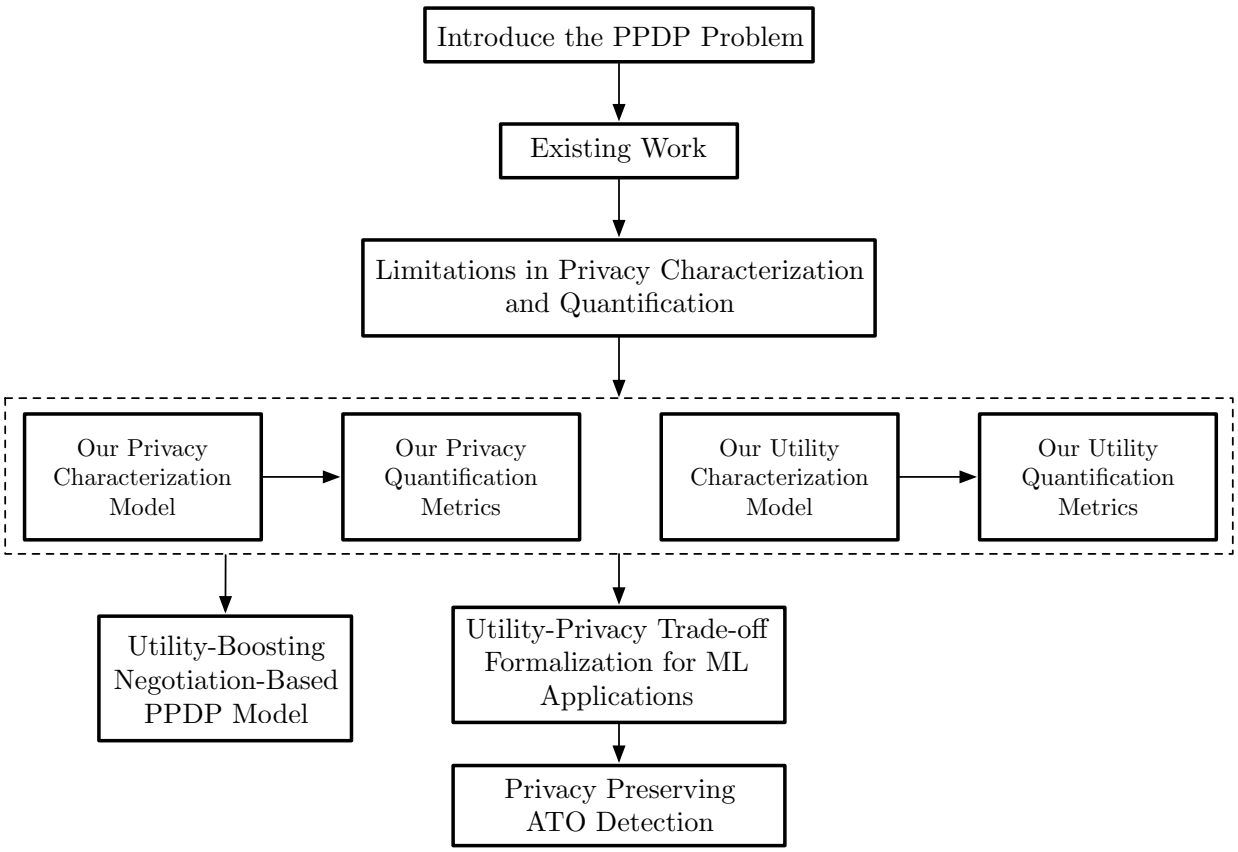


Figure 1.1: The thesis milestones

1.3.4 Privacy Preserving ATO Detection

Furthermore, we apply the proposed framework on a large-scale employee dataset from Barracuda Networks, a leader cybersecurity company, to predict the fraudulence of email account logins and thus detect possible Account Takeover (ATO) attacks. The results express variations in models' accuracy in binary classification of logins when trained on different datasets that satisfy different privacy constraints. The proposed framework enables a data publisher to quantitatively manage the utility-privacy trade-off and provide deeper insights about the value of the released data as well as the underlying privacy losses.

1.4 Thesis Organization

The rest of the thesis is structured as follows. In Chapter 2, we present the proposed privacy characterization. The proposed privacy quantification metrics are provided in Chapter 3. In Chapter 4, we introduce our utility loss metrics that enable us to formalize the utility-privacy tradeoff problem based on our data publishing model and propose a utility-boosting negotiation-based PPDP model. The utility-privacy tradeoff management framework for machine learning applications and a privacy preserving ATO detection model are proposed in Chapter 5. Finally, the thesis is concluded and future work is provided in Chapter 6.

Chapter 2

The Proposed Publishing Model and Privacy Characterization

2.1 Introduction

All previous approaches to characterize and quantify privacy have only investigated the privacy risks of publishing a sensitive attribute by focusing only on the change of belief of an adversary about the probability distribution of this attribute. However, we believe that any attribute by itself is not sensitive. The sensitivity of an attribute comes from combining it with other attributes. For example, cancer in a medical records dataset, and high or low salaries in an employees dataset, are not sensitive unless they are linked to a certain geographical area, age-range, or race. To obtain a meaningful definition of data privacy, it is necessary to characterize and quantify the knowledge about sensitive attributes that the adversary gains from observing the published dataset taking into consideration the combinational relation of different attributes. In our approach to characterize privacy, we employ a multi-dimensional scheme of privacy risk analysis attached with combining different attributes. Thus, in this chapter, after providing some essential preliminaries in Section 2.2, we introduce our combinational characterization of privacy in Sections 2.3, 2.4, and 2.5.

2.2 Preliminaries

In this section, we provide some technical background about the privacy preserving data publishing. We start by introducing the data publishing scenarios and explain how data is generally published. We then discuss different types of attacks on published datasets. Finally, a detailed analysis of the state-of-the-art existing PPDP techniques is presented.

2.2.1 Data Publishing

Privacy-Preserving Data Publishing Datasets publishing naturally consists of two phases. Different parties first collect data from record owners in a phase known as the data collection phase. It is then managed by the data publisher and released in a phase known as the data publishing phase. This data is published to a certain data recipient for the purpose of data mining or to the public for the purpose of providing useful societal information that could be utilized in different areas including research.

Data is commonly published in two models, untrusted and trusted model. In the untrusted model, the data publisher attempts to extract or manipulate sensitive information about record owners. To avoid such attempts, record owners apply cryptographic operations on the published data to prevent the publisher from accessing sensitive information. In the trusted model, the data publisher is assumed to be honest. In this model, record owners are not concerned about exposing their records, including the sensitive information, to the publisher. However, when data is released to the public, the publisher guarantees that sensitive information or identity of the record owner is not revealed to any possible adversary.

Utility-privacy Tradeoff Data utility is in a natural conflict with data privacy. It is trivial that, from the perspective of data utility, it is best to publish a dataset as is, while

from the perspective of data privacy, it is best to publish a mostly generalized dataset or even an empty one. Although this is easy to understand, as far as we know, including the information theoretic approaches proposed in [25] and [35], there is not yet a tight closed form relationship that fully model the utility-privacy trade off. We believe that the first step on the track of finding such a relationship is to better characterize and quantify both sides of the trade off. We note that the importance of studying data utility is undeniable and of great value as it definitely contributes to resolving the tradeoff modeling. In this thesis, we focus on providing more precise and practical approaches to quantify both data privacy and data utility and, in turn, lay out a reliable modeling of the tradeoff.

Data Disclosure Model Data is usually released in the format of tables, where the rows are the records of individuals and columns are their corresponding attributes. Some of the attributes are for information only and are not sensitive, while other attributes are individual sensitive attributes. For the information that is not being viewed as sensitive, when multiple records or maybe side information are combined, the individual maybe potentially identified. These attributes are generally referred to as *quasi-identifiers*, which may include information such as $\mathsf{Zip-Code}$, Age, and Gender. The sensitive information may include attributes that can uniquely identify the individuals such as the social security or the driving license numbers. These attributes are called *explicit-identifiers*. Another type of information being considered sensitive may include information such as $\mathsf{Disease}$ and Salary. When datasets are published, all explicit-identifiers are removed. Sensitive attribute disclosure occurs when the adversary learns information about an individual's sensitive attribute(s). This form of privacy breach is different and incomparable to learning whether an individual is included in a database, which is the focus of differential privacy [36].

Differential Privacy Since differential privacy has been grabbing a lot of attention

recently in the data privacy world, we would like to emphasize more on the irrelevance of our work to differential privacy efforts. Differential privacy is a classic privacy notion that has been widely investigated in literature. Although differential privacy has been attracting a lot of attention in the research community, there is a fundamental difference between the differential privacy and the syntactic privacy models such as k-anonymity, ℓ-diversity, t-closeness, as well as our proposed work, which focus on protecting sensitive attribute disclosure, while the goal of differential privacy is membership privacy.

In PPDP, once the data is published, it is available for any type of analysis, which is mainly targeted by syntactic privacy models. A typical scenario of PPDP is the impatient data release for public research purposes. A hospital possesses the data and is responsible for the privacy of patients participating in the dataset. The hospital's goal is to publish privacy preserving data regardless what kind of analysis or querying will be later applied to it.

Differential privacy models, on the other hand, typically target Privacy Preserving Data **Mining** (PPDM) [37–40]. In PPDM models, a data user aims at performing some data mining task on a set of private databases owned by different parties. The general idea of PPDM is to allow data mining from a modified version of the data that contains no sensitive information. In PPDM, as opposed to PPDP, the query that needs to be answered must be known prior to applying the privacy-preserving techniques. In the typical PPDM scenario, the data owner maintains control over the data and does not publish it. Instead, the owner responds to queries on the data, and ensures that the answers provided do not violate the privacy of the data subjects. In differential privacy, this is typically achieved by adding noise to the data, and it is necessary to know the analysis to be performed in advance in order to adjust the level of added noise. This approach contradicts with the main objective of data

publishing. In contrast, PPDP techniques ought to make the published data less precise than the original data but semantically truthful and hence preserve the integrity of the data.

In conclusion, syntactic privacy models are suitable for PPDP while differential privacy models are suitable for PPDM.

Generalization and Anonymization As the original dataset contains abundant information that could help an adversary link records to certain individuals, datasets are not published before being modified. Modifications could be accomplished in many ways. In general, all modifications are listed under the anonymization operations. These operations might be in the form of generalization, suppression, anatomization, permutation, or perturbation. Values of quasi-identifiers are somehow relaxed in case of generalization, or suppressed in case of suppression, to increase the range of individuals that carry the same quasi-identifier values and therefore increase the uncertainty of a possible adversary about certain individual's record [41–48]. On the other hand, anatomization and permutation operations achieve anonymization by dissociation of quasi-identifiers and sensitive attributes [49–52]. Perturbation mainly adds some noise to the whole dataset based on the statistical properties of the original data [53–57].

However, unlike statistical databases [50, 58], publishing individuals' data, also known as micro-data, requires that data remains intact after being released. Therefore not all the previously mentioned techniques are good candidates for anonymization of micro-data. To keep data intact, and as much useful as possible, it is obvious that, in most scenarios, only generalization and suppression operations could be applied in privacy-preserving micro-data publishing techniques.

2.2.2 Attacks on Datasets

Generally, there are two types of attacks on datasets, *record linkage* and *attribute linkage*. The record linkage occurs when some values of quasi-identifier attributes can lead to the identification of a smaller number of records in the published dataset. In this case, an individual having these attribute values is vulnerable to being linked to a limited number of records. On the other hand, attribute linkage occurs if some sensitive values are predominate in a group, where an attacker has no difficulty to infer such sensitive values for the record owner belonging to this group.

Attribute linkage mainly consists of two types, *homogeneity and background knowledge attacks*. In homogeneity attacks, anonymization model may create groups that leak information due to lack of diversity in the sensitive attribute. In fact, some anonymization process is based on generalizing the quasi-identifiers but does not address the sensitive attributes that can reveal information to an attacker. In background knowledge attacks, an attacker can have prior knowledge that enables him to guess sensitive data with high confidence. These kinds of attacks depend on other information available to an attacker. Using this background knowledge, an adversary can disclose information in two ways, *positive and negative disclosure*. In positive disclosure, an adversary can correctly identify the value of a sensitive attribute with high probability. On the other hand, in negative disclosure, the adversary can correctly eliminate some possible values of sensitive attribute with high probability. We also note that a background knowledge attack is difficult to prevent as compared to homogeneity attack.

In the next subsection we introduce a thorough analysis of the existing privacy-preserving data publishing techniques that attempt to combat these types of attacks on privacy.

2.2.3 Analysis of the Existing PPDP Schemes

In order to protect data privacy from different attacks, many privacy-preserving techniques and strategies have been proposed to meet different individual's privacy and utility requirements. All privacy-preserving techniques typically aim at protecting individual privacy, with minimizing impact on published data utility. k-anonymity, ℓ-diversity, and t-closeness are the techniques of most interest. In this section, to get an idea of how they overcome different attacks, we provide an overview of these privacy-preserving data publishing techniques. We discuss the design models for each technique, provide examples of how and to which extent privacy is achieved, and analyze their limitations.

2.2.3.1 k-anonymity

A table satisfies k-anonymity if every record in the table is indistinguishable from at least $k-1$ other records with respect to every set of *quasi-identifier* QID attributes; such a table is called a k-anonymous table. To satisfy this condition, the original table is generalized before being published. The generalized table forms groups with combinations of values of quasi-identifiers. Each group is named as an *equivalence class* $[C]$ and individuals within this class share the same combination of quasi-identifiers. Hence, for each combination of these values of the quasi-identifiers in the k-anonymous table, there are at least k records that share those values.

The idea of k-anonymity was proposed to combat record linkage attacks. An adversary who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than $1/k$. In [49, 59, 60], authors show that k-anonymity adds some protection against record linkage where it restricts the record linkage threats to a certain level. However it does not provide sufficient protection against

Table 2.1: Original table

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	49	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table 2.2: A 3-anonymous version

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

attribute linkage. k-anonymity is proved to be vulnerable against the homogeneity attack and the background knowledge attack.

Example 1 (Homogeneity and Background Knowledge Attacks)**.** Table 2.1 represents the original data table and Table 2.2 is an anonymized version of it satisfying 3-anonymity. The $\mathsf{Disease}$ attribute is sensitive. Suppose Alice knows that Bob is a 27-year old man living in $\mathsf{Zip-Code} = \mathsf{47678}$ and Bob's record is in the table. From Table 2.2, Alice can conclude that Bob is the owner of one of the first three records, and thus, must have $\mathsf{Heart-Disease}$. This is the homogeneity attack. For an example of the background knowledge attack, suppose that by knowing Carl's Age and $\mathsf{Zip-Code}$, Alice can conclude that Carl corresponds to a record in the last equivalence class in Table 2.2. Furthermore, suppose that Alice knows

that Carl has a very low risk for $\mathsf{Heart} - \mathsf{Disease}$. This background knowledge enables Alice to conclude that Carl most likely has Cancer.

To address the limitations of k-anonymity, [59] introduced ℓ-diversity as a stronger notion of privacy.

2.2.3.2 ℓ-diversity

An equivalence class is said to have ℓ-diversity if there are at least ℓ *well-represented* values for the sensitive attribute. A table is said to have ℓ-diversity if every equivalence class of the table has ℓ-diversity. [59] gave a number of interpretations of the term *well-represented*. The simplest understanding of *well-represented* would be to ensure that there are at least ℓ distinct values for the sensitive attributes in each equivalence class.

ℓ-diversity represents an important step beyond k-anonymity in protecting against attribute linkage. However it has several limitations and vulnerabilities to different attacks. First of all, distinct ℓ-diversity does not prevent probabilistic inference attacks. An equivalence class $[C]$ may have one attribute value that appears more frequent than other values. This enables an adversary to conclude that an individual u in $[C]$ is very likely to have that value. Therefore, [59] introduced more restrictive versions of ℓ-diversity such as entropy ℓ-diversity and recursive (c, ℓ)-diversity. These versions add more constraints on the published dataset. Thus, depending on the original dataset, the published dataset that satisfies these constraints may not always be achievable. Moreover ℓ-diversity is susceptible to attacks such as skewness and similarity attacks. We now briefly introduce these two attacks on ℓ-diversity using examples from [5].

When the overall distribution is skewed, satisfying the ℓ-diversity does not prevent attribute linkage. Consider the following example:

Example 2 (Skewness Attack)**.** Suppose that the original dataset has only one sensitive attribute, which is the test result for a particular virus. The virus takes two values either positive or negative. For a table that has 10,000 records, with 99% of them being negative and only 1% being positive. To satisfy distinct 2-diversity, any equivalence class $[C]$ must carry the two attribute values. If one of the equivalence classes has an equal number of positive and negative records, although it is 2-diverse, it presents a serious privacy risk. Any individual in this class has probability 50% to be infected compared to a 1% of the whole original population. Now, consider another extreme case. An equivalence class that has 49 positive records and only 1 negative record. Any individual in the equivalence class is 98% positive, compared to 1% of the whole original population.

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. Consider the following example:

Example 3 (Similarity Attack)**.** In the original Table 2.3 and an anonymized version satisfying distinct 3-diversity Table 2.4, consider Salary and Disease as the two sensitive attributes. An adversary is interested in finding the sensitive attribute value of an individual u. Based on the quasi-identifier values of u, an adversary is able to determine that the individual belongs to the first equivalence class. Therefore they know that their salary is in the range $[3K, 5K]$. This also applies to categorical attributes such as the Disease. The adversary would also know that the individual of interest indeed has a stomach-related disease. This loss of sensitive information occurs because ℓ-diversity does not take into account the semantic closeness of attribute values.

To prevent such attacks, authors in [5] proposed a privacy model, known as t-closeness.

Table 2.3: Original dataset

	Zip Code	**Age**	**Salary**	**Disease**
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 2.4: A 3-diverse version

	Zip Code	**Age**	**Salary**	**Disease**
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 2.5: 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

	Zip Code	**Age**	**Salary**	**Disease**
1	4767*	≤ 40	3K	gastric ulcer
2	4767*	≤ 40	5K	stomach cancer
3	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	4760*	≤ 40	4K	gastritis
8	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

2.2.3.3 t-closeness

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. The distance used in this publishing technique is the Earth Mover's Distance (EMD). EMD is simply the minimal amount of work needed to transform one distribution to another by moving distribution mass between each of them. Table 2.5 shows another anonymized version of Table 2.3 that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease. The similarity attack is prevented in Table 2.5 where, Revisiting Example 3, Alice can neither infer that Bob has a low salary nor he has a stomach-related disease.

As we previously mentioned, in t-closeness the value of t lacks any practical interpretation. In other words, we can hardly find any relation between t and the privacy loss. In k-anonymity and ℓ-diversity, given a k or ℓ value, the data publisher will have some intuition on its practical meaning in the real application and hence can effectively choose k and ℓ values to process the dataset. Unlike k-anonymity and ℓ-diversity, in t-closeness, the value t is merely an abstract distance between two distributions, that could have different meanings in different contexts. t-closeness also has several other limitations and weaknesses

[61]. First, it does not offer the flexibility of having different protection levels for different sensitive attribute values. Second, the EMD function, used to measure the distance between distributions, is not suitable for protection against attribute linkage on numerical sensitive attributes [62]. Third, as the case for ℓ-diversity, enforcing t-closeness would greatly affect the data utility where it requires the distribution of sensitive attribute values to be the same in all q equivalence classes. This would significantly damage the correlation between the set of *quasi-identifiers* QID and sensitive attributes. Finally, and the most important, we believe that the distance t measured as the EMD is unreliable to quantify the amount of privacy loss. More specifically, if we have two published tables T_1' and T_2' with $t_1 < t_2$, then table T_1' is not necessarily more privacy-preserving than T_2'. In other words, two published classes might have the same EMD distance relative to an original distribution, however they correspond to different levels of privacy loss. Consider the following example:

Example 4. A medical dataset has the Disease as a sensitive attribute. Distribution of attribute values *Cancer*, *Heart-Disease* and *Flu* in the original table is $(0.1, 0.5, 0.4)$. The published table is divided into two equivalence classes, denoted as $[C_1]$ and $[C_2]$. In $[C_1]$, distribution of attribute values is given as $(0.2, 0.4, 0.4)$, while in $[C_2]$ the distribution is $(0, 0.6, 0.4)$. This table achieves an 0.1-closeness w.r.t. Disease. Although the EMD in the two equivalence classes is the same, it is obvious that attribute values of individuals in $[C_2]$ are more prone to be inferred.

In section 3.4, supported by an analyzed example on t-closeness, we show how our proposed privacy metrics enable us to deliberately characterize and quantify this loss.

2.3 Generalization Model

In our privacy characterization we assume that any individual in a given table T only owns one record. Thus we, interchangeably, use the notion u to represent the record or the record owner. Let $U = \{u_n\}_{n=1}^{N}$ be N individuals participating in the data table T, $\mathcal{A} = \{A_l\}_{l=1}^{L}$ be the set of L attributes, and $u_n[A_l]$ be the value of attribute A_l for individual u_n. We denote the set of *quasi-identifiers* as $\mathsf{QID} \subset \mathcal{A}$. We also assume that any record is represented as a function of multi-variables $V = \{v_l\}_{l=1}^{L}$, where V corresponds to the set of attributes $\mathcal{A} = \{A_l\}_{l=1}^{L}$ in the original dataset. The order of each variable v_l, denoted as $\text{ord}(v_l) = |v_l|$, is the number of all possible attribute values.

To satisfy the privacy constraints, data disclosure techniques apply some generalizations to the quasi-identifiers $\mathsf{QID}s$ to avoid linking individuals to records in the table. Any value in the original table is mapped to a generalized value in the disclosed table following a certain mapping function. Records are generalized and represented as functions of multi-variables $V' = \{v'_l\}_{l=1}^{L}$, where V' is the generalization of V. The order of each generalized variable v'_l is defined as $\mathsf{ord}(v'_l) = \left|v'_l\right|$. After generalization, different combinations of v'_l's in the disclosed table T' naturally divide the table into a set $\mathcal{C} = \{[C_q]\}_{q=1}^{Q}$ of Q equivalence classes.

Consider two tables (T, T'), their corresponding attributes (V, V') and a mapping function $f: T \to T'$. We define table generalization as follows:

Definition 1 (**Table Generalization**)**.** For (T, T') and (V, V'), table generalization is a mapping $f: T \to T'$ that maps any table T to a table T'. This mapping function implies the following properties

- **Value Mapping**: $\forall v_l \in T$ and $v'_l \in T'$, any value $u[v_l]$ in T is mapped to $u'[v'_l]$ in T'.
- **Record Mapping**: For the two sets $V = \{v_1, v_2, \cdots, v_L\} \in T$ and $V' = \{v'_1, v'_2, \cdots, v'_L\} \in$

T', any record $u[V]$ in T is mapped to $u'[V']$ in T'.

- For any variable v_l and its generalization v'_l, it always holds that $\mathsf{ord}(v_l) \geq \mathsf{ord}(v'_l)$.
- After generalization, different combinations of v'_l's in the published Table T' naturally divide the table into a set $\mathcal{C} = \{[C_1], [C_2], \cdots, [C_Q]\}$ of Q equivalence classes.

Table generalization, represented in the mapping function, is the tool that controls privacy level of individuals and data utility of the published dataset. This mapping function is the key for designing any data publishing technique. Furthermore, privacy loss is directly linked to the combination of different variables. Hence, any publishing technique should consider the privacy risk attached with the combination of any of these variables.

Publishing a table T' gives different privacy risks for each combination of the generalized variables $< v'_i, v'_j >$. For example, $< \mathsf{Age}, \mathsf{Disease} >$ mapped to $< v'_1, v'_2 >$ is a combination of two variables that represents privacy risk of individuals of specific *Age* (age-range) and suffering from a specific $\mathsf{Disease}$, while $< \mathsf{Zip-Code}, \mathsf{Salary} >$ mapped to $< v'_3, v'_4 >$ is a combination that represents privacy risk of individuals living at a certain geographical area and are paid certain salary. Similarly, $< \mathsf{Age}, \mathsf{Zip-Code}, \mathsf{Salary}, \mathsf{Disease} >$ mapped to $< v'_1, v'_3, v'_4, v'_2 >$ represents the risk of individuals with certain *Age* v'_1, living at certain location with $\mathsf{Zip-Code}$ v'_3, suffering from $\mathsf{Disease}$ v'_2 and are paid an annual Salary v'_4. As the number of combined variables increases, the privacy risk of an individual increases and it would be easier for an adversary to identify an individual of interest from the published table. The order of any combination of variables could be easily derived as $\prod_{i=1}^{t} |v'_i|$, where t is the number of combined variables.

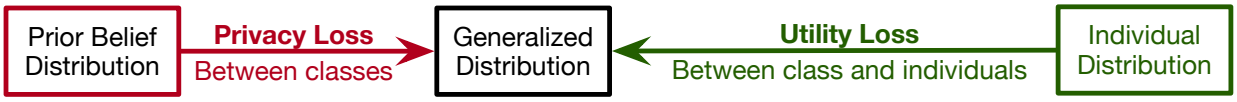


Figure 2.1: Our privacy and utility characterization approach

2.4 The Adversarial Prior and Posterior Beliefs

The adversary is given the published table T' generated from an original table T, and assumed to know quasi-identifier values $u[\mathsf{QID}]$ of an individual u of interest. The individual of interest is assumed to be in the table with probability 1. Hence, the membership disclosure problem, i.e., learning whether a given individual is present in the published dataset, is a different, incomparable privacy property and is out of the scope of this thesis.

2.4.1 The Adversarial Prior Belief

In our approach of characterizing privacy, shown in Fig. 2.1, an adversary is generally assumed to be aware of all the public information that might be available. Therefore, an adversary is believed to possess the original distribution of all the attributes. Moreover, for a dataset with L attributes, while some attributes are entirely independent, others could be correlated. Thus, an adversary possibly has an estimate of the joint distributions of these attributes. We now introduce the definition of the adversarial prior belief, that is the general public belief of all the distributions of attributes combinations.

Definition 2 (**Adversarial Prior Belief**)**.** For the set of attributes $\mathcal{A} = \{A_1, A_2, \cdots, A_L\}$ mapped to variables $V = \{v_1, v_2, \cdots, v_L\}$, an adversarial prior belief is modeled as

Original Distribution of Attributes: $\forall v_i \in V$, the original distribution of any random variable v_l given as a_{v_l} is previously known by an adversary.

Estimated Conditional Distribution of Attributes: $\forall v_l \in V$, an estimate of the

conditional distribution a_{v_i,v_j} of any combination of random variables is previously known by an adversary and is defined as

$$a_{v_i,v_j} = \tilde{P}(v_i \mid v_j),\ i,j = 1,\cdots,L,$$

where $\tilde{P}(v_i \mid v_j) = \frac{\tilde{P}(v_i \cap v_j)}{P(v_j)}$ and $\tilde{P}(v_i \cap v_j)$ is the estimated joint probability of any two attribute values.

For example, the distribution of a population over attributes such as Gender, Age and Disease is publicly available. Typically, within any geographical location, information such as percentage of males and females, percentages of individuals lying in a specific age-range and percentage of population suffering from a specific disease, are considered as adversarial prior information. Moreover, based on general trivial information, an adversary could have a very good estimate of joint distributions of some attributes. For instance, individuals suffering from a disease such as Breast Cancer are generally much more likely to be females, while individuals suffering from diseases such as Alzheimer and Arthritis are more likely to be above 60. Similarly, individuals living in a richer neighborhood are more likely to be paid higher salaries.

2.4.2 The Adversarial Posterior Belief

We believe that any adversarial model should take such information into consideration. Consequently, any privacy quantification approach that ignores this adversarial knowledge is not precise and lacks sufficiency. Any further information gained by an adversary after observing a published table is considered privacy loss and is represented as the adversarial posterior belief and is defined as follows,

Definition 3 (**Adversarial Posterior Belief**)**.** In a published table T', for the set of attributes $\mathcal{A} = \{A_1, A_2, \cdots, A_L\}$ mapped to variables $V' = \{v'_1, v'_2, \cdots, v'_L\}$, an adversarial posterior belief is modeled as

Published Conditional Distribution of Attributes: $\forall v'_i \in V$, the conditional distribution $x_{v'_i, v'_j}$ of any combination of random variables is defined as

$$x_{v'_i, v'_j} = P(v'_i \mid v'_j), i, j = 1, \cdots, L,$$

where $P(v'_i \mid v'_j) = \frac{P(v'_i \cap v'_j)}{P(v'_j)}$ and $P(v'_i \cap v'_j)$ is the published joint probability of any two attribute values.

As any published table T' is eventually formed of a subset of all possible combinations of generalized attributes v'_l, each of these combinations represents an equivalence class $[C_q]$.

While a_{v_i, v_j} and $x_{v'_i, v'_j}$ represent the prior and posterior beliefs of an adversary about an attribute v_i given an attribute v_j. We are generally interested in $a_{v_l, [C_q]}$ and $x_{v'_l, [C_q]}$, that are the prior and posterior beliefs of an adversary about an attribute v_l given a combination of generalized attributes represented in the specific class $[C_q]$ they form.

2.5 Defining Privacy Loss

The goal of any privacy-preserving technique is to minimize the privacy loss between prior and posterior belief as much as possible while maintaining a sufficient level of published data utility. As shown in Figure 2.2, we define this loss as the conditional privacy loss.

Definition 4 (**Conditional Privacy Loss**)**.** The privacy loss of an individual u belonging to an equivalence class $[C_q]$ with respect to an attribute v_l is the amount of information

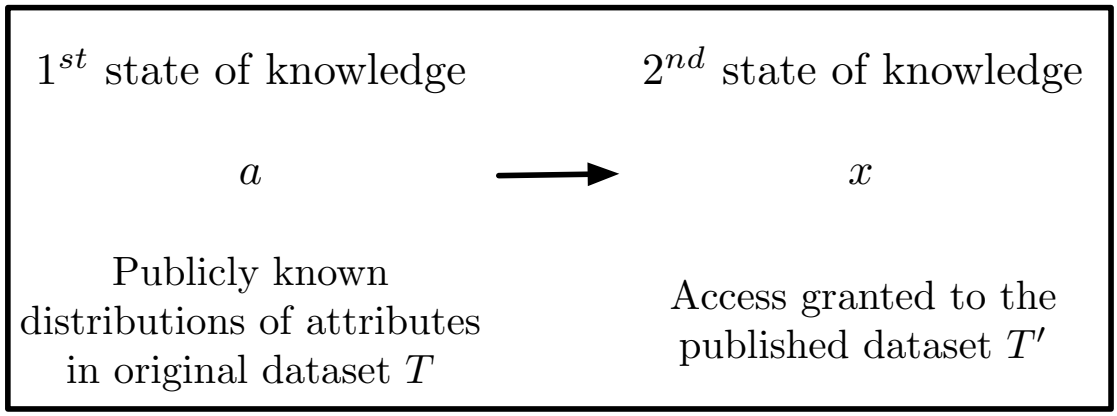


Figure 2.2: Definition of privacy loss

gained by an adversary represented as the change of the belief after publishing the table T'. This loss $L_P(v_l \,|\, [C_q])$ is typically the change of an adversarial belief about an attribute's distribution from $a_{v_l,[C_q]}$ to $x_{v_l,[C_q]}$.

Consider, an original table T having only 100 records described over two attributes, Age and Disease. If $P(5*) = \frac{1}{4}$, $P(Cancer) = \frac{1}{10}$ and an adversary has an estimate of their joint probability to be $\tilde{P}(\text{Cancer} \cap 5*) = \frac{4}{100}$, then the estimated conditional probability $\tilde{P}(\text{Cancer} \,|\, 5*) = \frac{\tilde{P}(\text{Cancer} \cap 5*)}{P(5*)} = \frac{16}{100}$. However, in the published table T', the adversary observes that the published joint probability $P(\text{Cancer} \cap 5*) = \frac{7}{100}$, which gives the published conditional probability $P(\text{Cancer} \,|\, 5*) = \frac{P(\text{Cancer} \cap 5*)}{P(5*)} = \frac{28}{100}$. Now an adversarial belief about individuals of the age-range (5∗) and suffering from Cancer has changed from a prior belief of 16/100 to a posterior belief of 28/100. This change of belief is the amount of information gained by an adversary. That is, the amount of privacy loss of individuals in a specific class (5∗) and having a certain attribute value (Cancer). Similarly we can find the privacy loss of individuals having other attribute values (other diseases) within the same class. One of our goals is to precisely quantify this loss. In the next chapter, we propose two privacy metrics that are able to measure privacy loss from two different perspectives.

We reemphasize that in our privacy quantification approach we only consider the loss between the two knowledge states. For example, let 10% of the individuals in a medical record's table T have *HIV*. If at the second state of knowledge, the adversary finds that an individual u_n has *HIV* with probability 10%, the information loss for this scheme should be 0 since this sensitive attribute's distribution is always considered public. Based on this, we introduce a generic definition of privacy-preserving data publishing as follows.

Definition 5 (**Privacy-Preserving Data Publishing**). Let $\mathcal{A} = \{A_1, A_2, \cdots, A_L\}$ be the set of all attributes. A published table T' is said to be privacy-preserving for set of individuals $U = \{u_1, u_2, \cdots, u_N\}$ if for any individual $u_n \in U$:

$$p(u_n[v_l]) = p(u_n[v'_l] \,|\, T'),\ n = 1, \cdots, N,\ l = 1, \cdots, L,$$

where each $u_n \in U$ represents an individual from the population, $p(u_n[v_l])$ denotes the probability of u_n on attribute v_l and $p(u_n[v'_l] \,|\, T')$ denotes the conditional probability of $u_n[v'_l]$ after the table T' is published.

In this definition, for a published table T' to be considered as privacy-preserving, the publishing technique strictly prohibits any privacy loss in the published data. While this conservative definition is practically impossible to achieve, any publishing technique should attempt to be as close as possible to achieve it. Privacy loss should, therefore, be quantified to be able to decide how far any given published data is from being privacy-preserving.

2.6 Summary

In this chapter, we introduced a comprehensive characterization of privacy to pave the way for solving the problem of privacy quantification in privacy-preserving data publishing. In order to consider the privacy loss of combined attributes, we presented data publishing as a multi-relational model. We re-defined the prior and posterior beliefs of the adversary. The proposed model and adversarial beliefs contribute to a more precise privacy characterization and quantification.

Chapter 3

The proposed Privacy Quantification Metrics

3.1 Introduction

There is an immense amount of existing privacy loss quantification metrics in literature [7]. The state-of-the-art approaches to measure privacy can be mainly sub-categorized into uncertainty, information gain or loss, similarity and diversity, and indistinguishability metrics [63, 64]. *Uncertainty metrics* measure the uncertainty in the adversarial estimate. The more uncertain the adversary is, the higher the achieved privacy in the published dataset. *Information gain or loss metrics* quantify the amount of information gained by the adversary, or the amount of information lost by users after data publishing. High adversarial gain and high user's loss of information corresponds to low privacy. *Similarity and diversity metrics* measure the similarity or diversity between the original and the published dataset. High similarity or low diversity between the two datasets corresponds to low privacy. *Indistinguishability* measures the ability of an adversary to distinguish between two outcomes of a privacy preserving data publishing technique. Privacy is high if it is hard for an adversary to distinguish between any pair of outcomes. Examples of such metrics are differential privacy, computational privacy, and distributed differential privacy [65].

In this chapter, we introduce our privacy loss quantification approach. Based on this approach we propose two different loss metrics. We explain the intuition behind these metrics, prove their correctness, and describe how they successfully contribute to the quantification of different privacy loss instances. We also define the threshold conditions to these metrics in order to be further used in the utility-privacy tradeoff problem formulation. Finally, we show the effectiveness of the proposed privacy characterization and quantification metrics through extensive empirical analysis and simulations results.

3.2 The Intuition Behind the Proposed Metrics

Our approach to quantify privacy mainly depends on understanding when information loss happens and how this loss could be measured. To have a better understanding of when loss occurs, we revisit the two states of knowledge of an adversary before and after a table T is published. At the first state of knowledge, based on public information of sensitive attribute's distribution, an adversary has some prior belief about the attribute value of an individual. This prior belief is in the form of probability distributions of attributes and joint distributions of their combinations. After publishing the table, an adversary moves to the second state of knowledge to gain some more information about the individual. This amount of information is the loss that we need to capture where it enables us to measure the extent to which this data publishing model minimizes privacy loss. We now analyze this loss and find a set of appropriate metrics that contribute to a better quantification of privacy represented in the amount of uncertainty an adversary has about an individual's sensitive attribute value after a table is published.

Before we introduce our two proposed privacy quantification metrics, namely, distribution

loss and entropy loss, we show the reason behind adopting these metrics.

The intuitive expectation of the proposed metrics is to compute the change in the data user's belief about an individual's sensitive attribute value before and after data disclosure. To find suitable metrics, we seek the distance measures based on two criteria. First is the *sensitivity* meaning that the metric should be sensitive to variations in the distributions. Second is the *independence* meaning that the metrics should be independent. That is, if two metrics independently measure the distance between two distributions, they both contribute to two different types of losses. As shown in Fig. 3.1, the L_1 and the Euclidean distances are the most sensitive metrics in comparison to others. However, the L_1 distance has the problem of not being robust under simple transformations such as rotation of the coordinate system. Therefore, it is not a good metric so we choose the Euclidean distance as our first distance metric. From Fig. 3.1 we can also see that entropy distance is the only metric that is independent of the other metrics which makes us speculate that the entropy loss metric will potentially account to some privacy loss instances that other metrics fail to represent. This qualifies it to be our second adopted metric.

We now introduce these two proposed privacy metrics and show how they are able to quantify the privacy loss from two different perspectives.

3.3 The Proposed Privacy Quantification Metrics

Let $S = \{s_i\}_{i=1}^{m}$ be the set of all m attribute values of a sensitive attribute $S \in \mathcal{A}$ (e.g. Disease in a medical dataset). The estimated initial distribution of S for equivalence class $[C_q]$ is given as $a_{S,[C_q]} = (a_1, a_2, \cdots, a_m)$. The disclosed distribution of S in an equivalence class $[C_q]$ is given as $x_{S,[C_q]} = (x_1, x_2, \cdots, x_m)$. Throughout the rest of this thesis, we

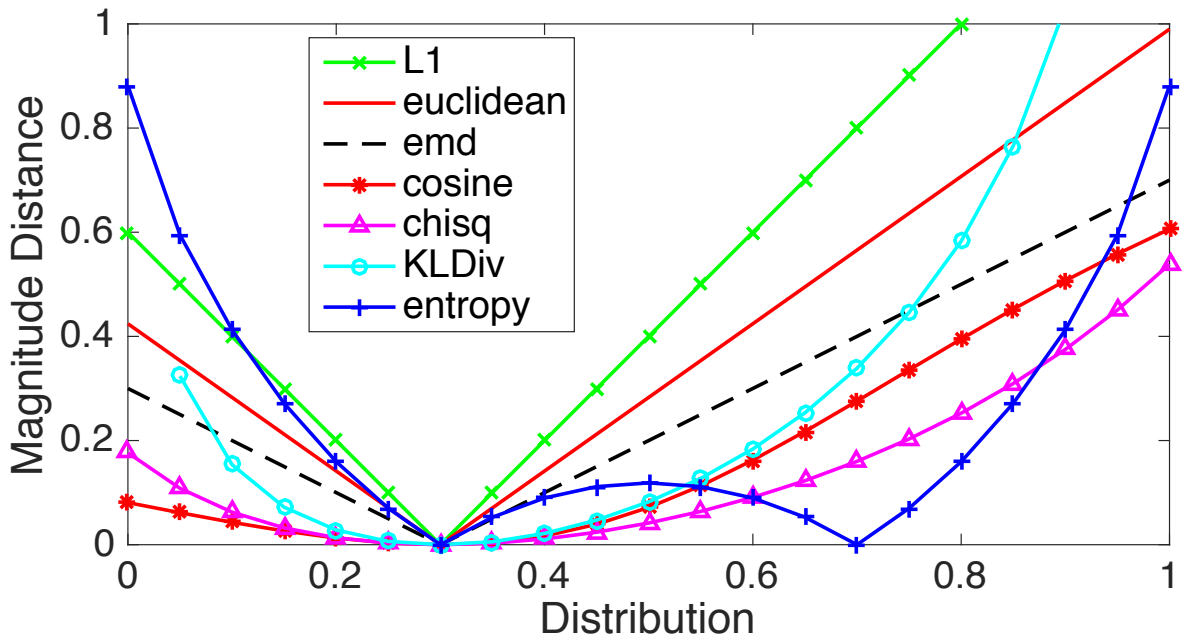


Figure 3.1: Comparison of different statistical distance metrics

denote $a_{S,[C_q]}$ as a and $x_{S,[C_q]}$ as x. We note that all of the concepts in this thesis are easily explained in the single sensitive attribute setting, but can also be extended to multiple sensitive attributes.

3.3.1 The Distribution Privacy Loss Metric

The distribution privacy loss could be viewed as a measure of the overall divergence of attribute values distribution from one state to the other. Generally, any privacy-preserving data publishing technique modifies the original dataset into a set of equivalence classes. The loss is measured between the original distribution of sensitive attribute values in the original and the published dataset for each given equivalence class. We, therefore, give the following definition.

Definition 6 (**Distribution Loss**)**.** For an individual u belonging to an equivalence class $[C_q]$, the distribution loss of attribute S given an equivalence class $[C_q]$ is defined as the

Euclidean distance between the two distributions a and x

$$\mathscr{L}_{P_D}(S, [C_q]) = \sqrt{\sum_{i=1}^{m} (a_i - x_i)^2}.$$

Since it is a Euclidean distance function, the distribution loss $\mathscr{L}_{P_D}(S, [C_q])$ defined above is indeed a distance metric, i.e. it satisfies all metric conditions.

As some privacy-preserving publishing techniques, falsely, assume that a uniform published distribution of attribute values achieves optimal privacy. It is interesting to find the distribution loss for this specific scenario. We find that the distribution loss is closely related to the standard deviation of the original distribution a, and the number of attribute values m in the sensitive attribute S.

Theorem 1. Let $S = \{s_1, s_2, \cdots, s_m\}$ be the set of all sensitive attribute values of a given dataset and $a = (a_1, a_2, \cdots, a_m)$ is the corresponding probability distribution. An individual u, belonging to an equivalence class $[C_q]$, has probability distribution on S of $x = (x_1, x_2, \cdots, x_m)$. The distribution loss of an attribute S in the published table T' with respect to uniform distribution is

$$\mathscr{L}_{P_D}(S, [C_q]) = \sqrt{\sum_{i=1}^{m} \left(a_i - \frac{1}{m}\right)^2} = \sigma_a \sqrt{m},$$

where σ_a is the standard deviation of a.

Proof. While the distribution loss is given as

$$\mathscr{L}_{P_D}(S, [C_q]) = \sqrt{\sum_{i=1}^{m} \left(a_i - \frac{1}{m}\right)^2},$$

and

$$\sigma_a = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(a_i - \mu_a)^2},$$

where μ_a is the mean of the distribution a. Since $\mu_a = \frac{\sum_{i=1}^{m} a_i}{m}$ then,

$$\sigma_a = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(a_i - \frac{\sum_{i=1}^{m} a_i}{m}\right)^2}.$$

Since $\sum_{i=1}^{m} a_i = 1$, then,

$$\sigma_a = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(a_i - \frac{1}{m}\right)^2}.$$

Therefore we have

$$\mathscr{L}_{P_D}(S, [C_q]) = \sigma_a\sqrt{m}.$$

□

$\mathscr{L}_{P_D}(S, [C_q])$ reaches a minimum value of 0 when $a_i = \frac{1}{m}$. It reaches a maximum value of $\sqrt{\frac{m-1}{m}}$ when $a_i = 1$ for some attribute value i and 0 for other attribute values. It is obvious that the distribution loss depends on the standard deviation of the original distribution. Thus, a uniform distribution of published attribute values is not essentially optimal. We believe that matching the published distribution to the original estimated distribution would indeed achieve better privacy.

It is worth to study the effect of the size of m as it might be of concern that can lead to a curse-of-dimensionality as it grows. However, as the size of m corresponds to the number of attribute values of the sensitive attribute, in most practical scenarios, the value m will not be very large. If it happens to be high, the proposed metric will depend on the distribution

of the attribute values. For instance, in an extreme case, a published class might have a single attribute value. Assuming that this attribute value is the least represented in the original dataset, the distribution metric will still be approximately bounded by the $\sqrt{2}$. Also assuming that this attribute is uniformly distributed in the original dataset, the distance between the two distributions will be $\sqrt{\frac{m-1}{m}}$ which tends to 1 as m increases.

Considering the more interesting case, for an arbitrary original distribution a_i and any two arbitrary published class distributions x_i and y_i where $i = 1, \cdots, m$. As m increases it is natural that the privacy loss in the two classes will get closer. That is, $\text{Dist}(a_i, x_i)$ and $\text{Dist}(a_i, y_i)$ will tend to be small. This means that the privacy loss could be small measured based on this sensitive attribute, which is close to the reality. For example, if the attribute is age. It would be much easier to determine the right sensitive attribute value if there is only a few in the observed equivalence classes, such as every 10 years. When the range increases, it could be hard to determine the right sensitive attribute value, for example, on a yearly basis, or more difficult on a combined year and monthly basis since even if you know the age of someone, you may still get the month incorrect.

We believe that attempts of normalization will disorder our metric, rendering unequal privacy loss instances to be considered equal. In other words, normalization can lead to an unprecedented equalization between two extremely different privacy loss instances. For example, consider two datasets T_1 and T_2 with two sensitive attributes S_1 and S_2. S_1 has 5 different sensitive attribute values while S_2 has 50. Let $[C_1]$ and $[C_2]$ be two published classes, in the published tables T_1' and T_2', each with a single represented attribute value corresponding to S_1 and S_2 respectively. From $[C_1]$, an observer can infer that the sensitive attribute value of an individual of interest is 1 out of 5. However, in $[C_2]$, the observer infers that the sensitive attribute value is 1 out of 50 which preserves more privacy. A random

guess will have probability 20% and 2% to get the right equivalence class for these two cases respectively. Therefore, normalization in such case will render both instances with the same privacy loss.

When publishing a table T, it is optimum to maintain the same original distribution over the set of equivalence classes. That is, distribution loss $\mathscr{L}_{P_D}(S, [C])$ is desired to be zero. However it is natural that the distance between distributions will change. This change contributes to the privacy loss. Therefore, an objective of minimizing privacy loss is to keep the distribution loss among equivalence classes below a predetermined level.

Definition 7 (ε-**Distribution Loss**)**.** A published table T' is said to have an ε-distribution loss if it has distribution loss $\mathscr{L}_{P_D}(S, [C]) \leq \varepsilon$ for the set of all equivalence classes. That is

$$\max(\mathscr{L}_{P_D}(S, [C_q])) \leq \varepsilon, \;\; q = 1, 2, \cdots, Q.$$

3.3.2 The Entropy Privacy Loss Metric

Considering only one metric ignoring the effect of other potential metrics leads to insufficient privacy quantification. An intuitive example for this problem is reviewing a blood work. The medical status of a patient cannot be determined based on only one measure even if this particular measure is the most sensitive one. Instead, a physician has to review the relation between combinations of all measures in the blood work.

While the distribution loss captures the amount by which privacy of an attribute is leaked, it does not give a sufficient implication about privacy loss of individuals carrying different attribute values. Specifically, a small distribution loss in the published table might lead to a critical decrease in the amount of uncertainty of an adversary about the attribute value of a

certain individual of interest. This motivates us to think of an information theoretic metric that would capture this change of adversarial uncertainty before and after table publishing. Hence, we propose the following distance metric that will serve as our second privacy metric.

Definition 8 (**Entropy Distance**)**.** Let $S = \{s_1, s_2, \cdots, s_m\}$ be the discrete set of attribute values of a sensitive attribute, $\mathscr{A} = (a_1, a_2, \cdots, a_m)$ and $\mathscr{B} = (b_1, b_2, \cdots, b_m)$ be two probability distributions on S. The entropy distance between $\mathscr{A}$ and $\mathscr{B}$ is defined as the difference of the entropy of the two distributions. That is,

$$\mathscr{L}_{P_E}(\mathscr{A}, \mathscr{B}) \;=\; \left| \sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i} - \sum_{i=1}^{m} b_i \log_2 \frac{1}{b_i} \right|.$$

The entropy distance typically measures the difference of uncertainty of an adversary about the sensitive attribute value of an individual from one state to the other. We now give the following theorem about entropy distance.

Theorem 2 (***Triangle Inequality***)**.** For the proposed entropy distance, the triangle inequality holds true, that is

$$\mathscr{L}_{P_E}(\mathscr{A}, \mathscr{B}) + \mathscr{L}_{P_E}(\mathscr{B}, \mathscr{C}) \geq \mathscr{L}_{P_E}(\mathscr{A}, \mathscr{C}).$$

Proof. We split the proof into four cases.

Case 1: $\sum_{i=1}^{m} a_i \log_2 a_i \leq \sum_{i=1}^{m} b_i \log_2 b_i$ and $\sum_{i=1}^{m} b_i \log_2 b_i \leq \sum_{i=1}^{m} c_i \log_2 c_i$.

Then we have

$$\sum_{i=1}^{m} a_i \log_2 a_i \leq \sum_{i=1}^{m} c_i \log_2 c_i,$$

and

$$\begin{aligned}\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C}) &= -\sum_{i=1}^{m} a_i\log_2 a_i+\sum_{i=1}^{m} b_i\log_2 b_i-\sum_{i=1}^{m} b_i\log_2 b_i+\sum_{i=1}^{m} c_i\log_2 c_i\\ &= \left|-\sum_{i=1}^{m} a_i\log_2 a_i+\sum_{i=1}^{m} c_i\log_2 c_i\right|\\ &= \mathscr{L}_{P_E}(\mathscr{A},\mathscr{C}).\end{aligned}$$

Case 2: $\sum_{i=1}^{m} a_i\log_2 a_i \geq \sum_{i=1}^{m} b_i\log_2 b_i$ and $\sum_{i=1}^{m} b_i\log_2 b_i \geq \sum_{i=1}^{m} c_i\log_2 c_i$.

Then we have

$$\sum_{i=1}^{m} a_i\log_2 a_i \geq \sum_{i=1}^{m} c_i\log_2 c_i,$$

and

$$\begin{aligned}\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C}) &= \sum_{i=1}^{m} a_i\log_2 a_i-\sum_{i=1}^{m} b_i\log_2 b_i+\sum_{i=1}^{m} b_i\log_2 b_i-\sum_{i=1}^{m} c_i\log_2 c_i\\ &= \left|-\sum_{i=1}^{m} a_i\log_2 a_i+\sum_{i=1}^{m} c_i\log_2 c_i\right|\\ &= \mathscr{L}_{P_E}(\mathscr{A},\mathscr{C}).\end{aligned}$$

Case 3: $\sum_{i=1}^{m} a_i\log_2 a_i \leq \sum_{i=1}^{m} b_i\log_2 b_i$ and $\sum_{i=1}^{m} b_i\log_2 b_i \geq \sum_{i=1}^{m} c_i\log_2 c_i$.

Then we have

$$\prod_{i=1}^{m} a_i^{a_i} \leq \prod_{i=1}^{m} b_i^{b_i}, \text{ and } \prod_{i=1}^{m} b_i^{b_i} \geq \prod_{i-1}^{m} c_i^{c_i}.$$

and

$$\begin{aligned}\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C}) &= -\sum_{i=1}^{m} a_i\log_2 a_i+\sum_{i=1}^{m} b_i\log_2 b_i+\sum_{i=1}^{m} b_i\log_2 b_i-\sum_{i=1}^{m} c_i\log_2 c_i\\ &= \log\left(\prod_{i=1}^{m}\frac{b_i^{b_i}}{a_i^{a_i}}\cdot\frac{b_i^{b_i}}{c_i^{c_i}}\right)\geq\log\left(\prod_{i=1}^{m}\frac{b_i^{b_i}}{a_i^{a_i}}\right)\geq\log\left(\prod_{i=1}^{m}\frac{c_i^{c_i}}{a_i^{a_i}}\right)\\ &= -\sum_{i=1}^{m} a_i\log_2 a_i+\sum_{i=1}^{m} c_i\log_2 c_i.\end{aligned}$$

Similarly, we also have

$$\begin{aligned}\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C}) &\geq \log\left(\prod_{i=1}^{m}\frac{a_i^{a_i}}{c_i^{c_i}}\right)\\ &= \sum_{i=1}^{m} a_i\log_2 a_i-\sum_{i=1}^{m} c_i\log_2 c_i.\end{aligned}$$

Therefore, we have

$$\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C})\geq\mathscr{L}_{P_E}(\mathscr{A},\mathscr{C}).$$

Case 4: $\sum_{i=1}^{m} a_i\log_2 a_i\geq\sum_{i=1}^{m} b_i\log_2 b_i$ and $\sum_{i=1}^{m} b_i\log_2 b_i\leq\sum_{i=1}^{m} c_i\log_2 c_i$.

Then we have

$$\prod_{i=1}^{m} a_i^{a_i}\geq\prod_{i=1}^{m} b_i^{b_i},\ \text{and}\ \prod_{i=1}^{m} b_i^{b_i}\leq\prod_{i-1}^{m} c_i^{c_i}.$$

and

$$\begin{aligned}\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C}) &= \sum_{i=1}^{m} a_i\log_2 a_i-\sum_{i=1}^{m} b_i\log_2 b_i-\sum_{i=1}^{m} b_i\log_2 b_i+\sum_{i=1}^{m} c_i\log_2 c_i\\ &= \log\left(\prod_{i=1}^{m}\frac{a_i^{a_i}}{b_i^{b_i}}\cdot\frac{c_i^{c_i}}{b_i^{b_i}}\right)\geq\log\left(\prod_{i=1}^{m}\frac{a_i^{a_i}}{b_i^{b_i}}\right)\geq\log\left(\prod_{i=1}^{m}\frac{a_i^{a_i}}{c_i^{c_i}}\right).\\ &= \sum_{i=1}^{m} a_i\log_2 a_i-\sum_{i=1}^{m} c_i\log_2 c_i.\end{aligned}$$

Similarly, we also have

$$\begin{aligned}\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C}) &= \log\left(\prod_{i=1}^{m}\frac{a_i^{a_i}}{b_i^{b_i}}\cdot\frac{c_i^{c_i}}{b_i^{b_i}}\right)\\ &\geq \log\left(\prod_{i=1}^{m}\frac{c_i^{c_i}}{b_i^{b_i}}\right)\geq\log\left(\prod_{i=1}^{m}\frac{c_i^{c_i}}{a_i^{a_i}}\right)\\ &\geq -\sum_{i=1}^{m} a_i\log_2 a_i+\sum_{i=1}^{m} c_i\log_2 c_i.\end{aligned}$$

Therefore, we have

$$\mathscr{L}_{P_E}(\mathscr{A},\mathscr{B})+\mathscr{L}_{P_E}(\mathscr{B},\mathscr{C})\geq\mathscr{L}_{P_E}(\mathscr{A},\mathscr{C}).$$ □

Based on Theorem 2, we have the following theorem.

Theorem 3. Entropy loss is a distance metric and has the following properties:

1. **Non-negativity:** $\mathscr{L}_{P_E}(x,y)\geq 0$.

2. **Definiteness:** $\mathscr{L}_{P_E}(x,y)=0$ if and only if $x=y$

3. **Symmetry:** $\mathscr{L}_{P_E}(x,y)=\mathscr{L}_{P_E}(y,x)$.

4. **Triangle inequality:** $\mathscr{L}_{P_E}(x,z)\leq\mathscr{L}_{P_E}(x,y)+\mathscr{L}_{P_E}(y,z)$.

The proof of this theorem is straight forward. We note that maximum entropy of attribute values in the published dataset does not achieve the maximum privacy. While the maximum entropy corresponds to the uniform distribution of attribute values, this kind of distribution can be optimum if the background information of an adversary is ignored. However, given that an adversary has some prior belief about original attribute values distributions, it is best to maintain the same entropy level after publishing. Therefore, we introduce the following privacy metric.

Definition 9 (**Entropy Loss**)**.** For an individual u_n belonging to an equivalence class $[C_q]$, the entropy loss is defined as

$$\mathscr{L}_{P_E}(S, [C_q]) = \left| \sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i} - \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} \right|.$$

We define the entropy loss of an individual as the entropy loss of the equivalence class that the individual belongs to. Note that the entropy loss reaches maximum $\log_2 m$ when the original distribution is uniform and the published distribution is $x_i = 1$ for some attribute value i and 0 for other attribute values. This is easily explained as a transition in the adversarial belief, from a state where the adversarial uncertainty about the attribute value of an individual of interest in a given class is maximum, to a state where they become 100% confident about the attribute value of this individual. Given the knowledge of the distribution of sensitive attribute values of the original distribution, an adversary has a certain level of uncertainty about individuals attribute values. Any change in this level of uncertainty is considered a loss. An objective of any data publishing technique would be minimizing this loss.

Note that the entropy loss metric is convex in our case since $\sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i}$ is a fixed

number that can be computed based on the prior knowledge ahead of time. Therefore, the optimal value always theoretically exists.

Let $Z = \sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i}$, then for our application, Z is a fixed number that can be easily computed based on the prior knowledge ahead of time. The maximum value of

$$\left| \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} - Z \right|$$

is $\log_2 m - Z$. Then to minimize entropy loss is equivalent to minimize

$$\min_{x_i} \left(\sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} - Z \right), \text{ if } \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} \geq Z,$$

or

$$\max_{x_i} \left(\sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} - Z \right), \text{ if } \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} \leq Z.$$

For any Z, $\sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} - Z$ is convex. Therefore, the optimal value always exists theoretically. Moreover, for this metric, we also think that normalization will lead to an unprecedented equalization between two extremely different privacy loss instances as we have demonstrated in the previous metric.

As previously mentioned, many PPDP techniques assume that a uniform published distribution of attribute values achieves optimal privacy. Thus, we also find the entropy loss for this specific scenario. For a published uniform distribution x of a dataset, the entropy is $\log_2 m$. Using Definition 9, the entropy loss between the original dataset distribution and the uniform distribution is given as $\log_2 m - \sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i}$.

We also introduce a threshold condition on the entropy privacy loss metric which can be used by a DO to express the privacy loss expectations. This threshold condition will also be

later used in formalization and management of the utility-privacy trade-off.

Definition 10 (α-**Entropy Loss**)**.** A published table T' is said to have an α-entropy loss if it has entropy loss $\mathscr{L}_{P_E}(S,[C]) \leq \alpha$ for the set of all equivalence classes. That is

$$\max(\mathscr{L}_{P_E}(S,[C_q])) \leq \alpha, \;\; q = 1, 2, \cdots, Q.$$

We argue that distribution loss and entropy loss are two different metrics. To justify this argument, let us assume the case when the published attribute values' distribution, in a specific class, is a permutation of the original distribution. Unless the original distribution is uniform, whatever the distribution loss is, the entropy loss will always be zero. More examples are presented in the next section to support our argument.

Meanwhile, we do not know how many metrics would be sufficient to quantify privacy. However, we believe that any further proposed independent metrics that would contribute to reaching an optimum and provably sufficient set of measures, can be added to the proposed quantitative measurement framework.

3.4 Empirical Analysis and Simulation Results

This section is divided into two parts. In the first part, based on our findings, we introduce a wide set of empirical examples for different case scenarios that support our findings. The provided examples aim to help understand the implications of the proposed metrics and show how these metrics contribute to analyzing, comparing and evaluating the previously mentioned existing privacy-preserving data-publishing techniques. In the second part of this section, aided with our simulation results, we focus on instances where different PPDP

techniques assume to achieve an intended privacy level. However, based on our proposed metrics, they fail to express, and therefore fail to avoid, a considerable amount of privacy loss. Throughout this section, we assume that an adversary has no other side information about dataset statistics or the user of interest other than the determined quasi-identifier values.

3.4.1 Empirical Analysis

We begin by giving examples to show how the distribution and the entropy losses are two different measures of privacy loss.

Example 5. As demonstrated in Fig. 3.2 and Fig. 3.3, consider a dataset T with sensitive attribute S containing $m = 3$ attribute values. The original attribute values distribution of S is given as $\left(\frac{7}{12}, \frac{3}{12}, \frac{2}{12}\right)$. The published table T' is divided into a set of $q = 3$ equivalence classes with attribute values distributions of $\left(\frac{3}{4}, \frac{1}{4}, 0\right)$, $\left(\frac{3}{4}, \frac{1}{4}, 0\right)$ and $\left(\frac{1}{4}, \frac{1}{4}, \frac{2}{4}\right)$. The distribution loss $\mathscr{L}_{P_D}(S, [C])$ and the entropy loss $\mathscr{L}_{P_E}(S, [C])$ for the attribute values are $\left[\frac{\sqrt{8}}{12}, \frac{\sqrt{8}}{12}, \frac{\sqrt{26}}{12}\right]$ and $[0.57, 0.57, 0.11]$ respectively. We notice that $[C_3]$ has the highest distribution loss however it provides the least entropy loss. It is obvious that a high distribution loss does not necessarily provide a high entropy loss and vice versa. This motivates us to further think of the implication of the large distribution loss of $[C_3]$. The third attribute value is fully represented in this class. Therefore, an adversary has a 100% confidence that any individual that has the third attribute value is in $[C_3]$.

While a published table with uniform attribute values distribution naturally has the highest output entropy (not entropy loss), it is sometimes, falsely, assumed to be optimal. We believe that matching the published distributions with the original distribution would cer-

Original Distribution	Publishing Distribution
	$(\frac{3}{4}, \frac{1}{4}, 0)$
$(\frac{7}{12}, \frac{3}{12}, \frac{2}{12})$	$(\frac{3}{4}, \frac{1}{4}, 0)$
	$(\frac{1}{4}, \frac{1}{4}, \frac{2}{4})$

$$\mathscr{L}_{P_D}(S, \mathcal{C}) \text{ is } \left[\frac{\sqrt{8}}{12}, \frac{\sqrt{8}}{12}, \frac{\sqrt{26}}{12}\right]$$

$[C_3]$ has the highest $\mathscr{L}_{P_D}(S, \mathcal{C})$

Adversary has a 100% confidence that any individual in $[C_1]$ or $[C_2]$ doesn't have the third attribute value.

Figure 3.2: Example to show insufficiency of distribution loss for privacy quantification

Original Distribution	Publishing Distribution
	$(\frac{3}{4}, \frac{1}{4}, 0)$
$(\frac{7}{12}, \frac{3}{12}, \frac{2}{12})$	$(\frac{3}{4}, \frac{1}{4}, 0)$
	$(\frac{1}{4}, \frac{1}{4}, \frac{2}{4})$

$$\mathscr{L}_{P_D}(S, \mathcal{C}) \text{ is } \left[\frac{\sqrt{8}}{12}, \frac{\sqrt{8}}{12}, \frac{\sqrt{26}}{12}\right]$$

$$\mathscr{L}_{P_E}(S, \mathcal{C}) \text{ is } [0.57, 0.57, 0.11]$$

$[C_3]$ has the highest $\mathscr{L}_{P_D}(S, \mathcal{C})$ however it provides the least $\mathscr{L}_{P_E}(S, \mathcal{C})$

A high distribution leakage does not necessarily provide a high entropy leakage and vice-versa.

Figure 3.3: Example to show the distinction of the proposed privacy metrics

tainly achieve a better privacy protection. This can be justified using the following example.

Example 6. Consider a dataset T with sensitive attribute S containing $m = 4$ attribute values. The original attribute values distribution of S is given as $\left(\frac{10}{16}, \frac{2}{16}, \frac{2}{16}, \frac{2}{16}\right)$. The published table T' is divided into a set of $Q = 4$ equivalence classes with attribute values distributions of $\left(\frac{4}{16}, \frac{4}{16}, \frac{4}{16}, \frac{4}{16}\right)$, $\left(\frac{12}{16}, \frac{4}{16}, 0, 0\right)$, $\left(\frac{12}{16}, 0, \frac{4}{16}, 0\right)$, and $\left(\frac{12}{16}, 0, 0, \frac{4}{16}\right)$. The distribution loss $\mathscr{L}_{P_D}(S, [C])$ and the entropy loss $\mathscr{L}_{P_E}(S, [C])$ for the attribute values are $\left[\frac{\sqrt{48}}{16}, \frac{4}{16}, \frac{4}{16}, \frac{4}{16}\right]$ and $[0.45, 0.73, 0.73, 0.73]$ respectively. It is obvious that the first equivalence class $[C_1]$ has a uniform distribution, however it does not achieve the best distribution loss.

Another example that shows how a uniform distribution of published attribute values for an equivalence class might, at some cases, even give a large distance and a worse entropy loss than other distributions.

Example 7. Consider a dataset T with sensitive attribute S containing $m = 4$ attribute values. The original attribute values distribution of S is given as $\left(\frac{9}{16}, \frac{3}{16}, \frac{2}{16}, \frac{2}{16}\right)$. The published table T' is divided into a set of $Q = 4$ equivalence classes with attribute values distributions of $\left(\frac{4}{16}, \frac{4}{16}, \frac{4}{16}, \frac{4}{16}\right)$, $\left(\frac{8}{16}, \frac{4}{16}, \frac{4}{16}, 0\right)$, $\left(\frac{12}{16}, \frac{4}{16}, 0, 0\right)$, and $\left(\frac{12}{16}, 0, 0, \frac{4}{16}\right)$. The distribution loss $\mathscr{L}_{P_D}(S, [C])$ and the entropy loss $\mathscr{L}_{P_E}(S, [C])$ for the attribute values are $\left[\frac{\sqrt{34}}{16}, \frac{\sqrt{10}}{16}, \frac{\sqrt{18}}{16}, \frac{\sqrt{26}}{16}\right]$ and $[0.33, 0.16, 0.85, 0.85]$ respectively. We notice here that the first equivalence class having a uniform distribution has the highest distribution loss and does not achieve the best entropy loss.

Apparently, the distribution loss maybe considered as a reflection of the extent to which privacy of attribute values are leaked. On the other hand, the entropy loss is a reflection of the extent to which privacy of individuals within an equivalence class is leaked compared to

the initial entropy from the original distribution. We stress that both losses contribute to the overall privacy loss of individuals in the published dataset.

Based on the previous examples, it is obvious that while designing any publishing technique, to achieve a better level of privacy, the data publisher should not only consider the distribution loss but also the entropy loss. Now we use our proposed metrics to analyze some existing schemes.

Example 8. In the example from [4], the original impatient dataset is given in Table 3.1 and the 4-anonymous impatient dataset is given in Table 3.2. For these two tables, the probability distribution of the sensitive attribute, Disease, is $\left(\frac{3}{12}, \frac{4}{12}, \frac{5}{12}\right)$. In this case, the distribution loss $\mathscr{L}_{P_D}(S, [C])$ for individuals within the first, second and third equivalence classes is $[0.5137, 0.2357, 0.7619]$, while the entropy loss is $[0.5546, 0.0546, 1.5546]$ respectively. Our findings for Table 3.2 can be summarized as follows:

1. Patients under 30 have *Heart-Disease* or *Virus-Infection* with equal probability. The scheme provides distribution loss $\mathscr{L}_{P_D}(S, [C_1]) = 0.5137$ and entropy loss equals $\mathscr{L}_{P_E}(< 30) = 0.5546$.

2. For patients over 40, $\frac{1}{4}$ have *Cancer*, $\frac{1}{4}$ have *Heart-Disease* and $\frac{1}{2}$ have *Virus-Infection*. The scheme provides distribution loss $\mathscr{L}_{P_D}(S, [C_2]) = 0.2357$ and entropy loss is $\mathscr{L}_{P_E}(\geq 40) = 0.0546$.

3. Patients in their 30s, all have *Cancer*. The individual gets distribution loss $\mathscr{L}_{P_D}(S, [C_3]) = 0.7619$ and entropy loss $\mathscr{L}_{P_E}(30s) = 1.5546$.

Example 9. For the same original impatient dataset from last example given in Table 3.1, the 3-diverse impatient dataset is given in Table 3.3. For these two tables, the probability distribution of the sensitive attribute, Disease, is $\left(\frac{3}{12}, \frac{4}{12}, \frac{5}{12}\right)$. In this case, the distribution loss

Table 3.1: Original dataset

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Virus Infection
4	13053	23	American	Virus Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Virus Infection
8	14850	49	American	Virus Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Table 3.2: 4-anonymous impatient micro-data

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Virus Infection
4	130**	<30	*	Virus Infection
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Heart Disease
7	1485*	≥40	*	Virus Infection
8	1485*	≥40	*	Virus Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Table 3.3: 3-diverse impatient micro-data

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	1305*	≤40	*	Heart Disease
2	1305*	≤40	*	Virus Infection
3	1305*	≤40	*	Cancer
4	1305*	≤40	*	Cancer
5	1485*	>40	*	Cancer
6	1485*	>40	*	Heart Disease
7	1485*	>40	*	Virus Infection
8	1485*	>40	*	Virus Infection
9	1306*	≤40	*	Heart Disease
10	1306*	≤40	*	Virus Infection
11	1306*	≤40	*	Cancer
12	1306*	≤40	*	Cancer

Table 3.4: Original dataset

	Non-Sensitive		Sensitive
	Zip Code	**Age**	**Disease**
1	49012	25	Flu
2	49013	28	Flu
3	49013	29	Heart Disease
4	49970	39	Flu
5	48823	49	Cancer
6	49971	34	Flu
7	48824	48	Heart Disease
8	48823	45	Cancer
9	48824	46	Flu
10	49971	37	Heart Disease
11	49012	22	Flu
12	49970	32	Flu

Table 3.5: 4-anonymous, 2-diverse dataset

	Non-Sensitive		Sensitive
	Zip Code	**Age**	**Disease**
1	4901*	2*	Flu
2	4901*	2*	Flu
3	4901*	2*	Flu
4	4901*	2*	Heart Disease
5	4997*	3*	Flu
6	4997*	3*	Flu
7	4997*	3*	Flu
8	4997*	3*	Heart Disease
9	4882*	4*	Flu
10	4882*	4*	Heart Disease
11	4882*	4*	Cancer
12	4882*	4*	Cancer

for individuals within the first, second and third equivalence classes is $[0.1179, 0.2357, 0.1179]$, while the entropy loss is $[0.0546, 0.0546, 0.0546]$ respectively. Our findings for Table 3.3 can be summarized as follows:

1. Patients under 40 and living in *Zip-Code = 1305** have *Heart-Disease*, *Virus-Infection* or *Cancer*. Therefore, the scheme provides distribution loss $\mathscr{L}_{P_D}(S, [C_1]) = 0.1179$ and entropy loss $\mathscr{L}_{P_E}(1305* \cap \leq 40) = 0.0546$.

2. For patients over 40 and living in *Zip-Code = 1485**, having *Heart-Disease*, *Virus-Infection* or *Cancer*. The scheme provides distribution loss $\mathscr{L}_{P_D}(S, [C_2]) = 0.2357$ and entropy loss $\mathscr{L}_{P_E}(1485* \cap > 40) = 0.0546$.

3. For patients under 40 and living in *Zip-Code = 1306**, having *Heart-Disease*, *Virus-Infection* or *Cancer*. The individual gets distribution loss $\mathscr{L}_{P_D}(S, [C_3]) = 0.1179$ and entropy loss $\mathscr{L}_{P_E}(1306* \cap \leq 40) = 0.0546$.

Example 10. In this example, the original impatient dataset is given in Table 3.4. The 4-anonymous, 2-diverse, 0.67-closeness impatient dataset is given in Table 3.5. For these two

tables, the original probability distribution for the three diseases is $\left(\frac{7}{12}, \frac{3}{12}, \frac{2}{12}\right)$. In this case, the EMD is $\left[\frac{1}{3}, \frac{1}{3}, \frac{2}{3}\right]$, distribution loss $\mathscr{L}_{P_D}(S, [C])$ for individuals within the first, second and third equivalence classes is $\left[\frac{\sqrt{8}}{12}, \frac{\sqrt{8}}{12}, \frac{\sqrt{26}}{12}\right]$, while the entropy loss is $[0.57, 0.57, 0.11]$, respectively.

We finally show how EMD in t-closeness is not reliable and insufficient to measure privacy loss.

For the original impatient dataset from [5] given in Table 2.5, an 0.167-closeness, w.r.t Salary, impatient dataset is given in Table 2.4. The values of the sensitive attribute Salary in the original table are $\{3K, 4K, 5K, 6K, 7K, 8K, 9K, 10K, 11K\}$. The values of the same attribute in the published table T' are given as $\{3K, 5K, 9K\}$, $\{6K, 11K, 8K\}$ and $\{4K, 7K, 10K\}$ for the three equivalence classes $[C_1]$, $[C_2]$ and $[C_3]$. In this case, the EMD for the three classes is given as $[0.167, 0.167, 0.083]$. The EMD, proposed in t-closeness, is a semantic metric. It gives a weight to the attribute values based on their sensitivity in the original distribution. However, as we will show, it fails to give a correct measurement of the privacy loss.

Example 11. Consider a 27 records dataset with same attribute values having the same uniform distribution. After publishing, this dataset is divided into 9 equivalence classes. We consider two possible equivalence classes, $[C_1]$ and $[C_2]$. Assuming that the sensitive attribute values in $[C_1]$ and $[C_2]$ are $\{3K, 4K, 5K\}$ and $\{7K, 7K, 7K\}$ respectively. The EMD for both classes is calculated as $[0.375, 0.278]$. Based on the EMD, $[C_2]$ achieves better privacy level among the two equivalence classes. However, it is obvious that the adversarial general belief about the attribute values before and after publishing has changed more dramatically in $[C_2]$ compared to $[C_1]$. This change of belief is properly characterized in the value of our distribution loss metric $\mathscr{L}_{P_D}(S, [C])$ which is given as $[0.22, 0.89]$. Furthermore, we can

easily notice that $[C_2]$ suffers from a higher privacy loss where all individuals have the same sensitive attribute value $(7K)$. Thus, an adversary would know the attribute value of an individual in this class with probability 1. This change of adversarial certainty about individual's attribute values is indeed a privacy loss. This loss is very well represented in our entropy loss metric $\mathscr{L}_{P_E}(S, [C])$ which is given as $[0.158, 3.16]$.

3.4.2 Simulation Results

In our simulations, we investigate the effectiveness of different PPDP techniques based on our privacy metrics. Simulation results give us a more insightful understanding of privacy loss. Specifically, our analysis gives a spotlight on several instances where published tables are believed to achieve privacy based on the utilized PPDP techniques, while based on our metrics, they do leak valuable private information about users in the datasets. We also show how our proposed metrics enable a data publisher to have more control over the privacy of a specific group of users having certain sensitive attribute values.

Simulations are done on a sample of the US census dataset from the UC Irvine machine learning repository [66]. After eliminating records with missing values, we have a total of 30,162 records. Following the work in [4], as shown in Table 3.6, we utilize only 9 attributes, 7 of which form the set of possible quasi-identifiers while Occupation and Salary form the set of possible sensitive attributes. We adopt the incognito algorithm [41] for generating the anonymized tables that satisfy the privacy measures of different PPDP techniques. Throughout the simulations, we consider the Occupation as the sensitive attribute. The number of quasi-identifiers QIDs is represented by the variable e that takes values from 1 to 7 with the same order in Table 3.6. While evaluating privacy of different PPDP techniques, it is essential to maintain the same level of data quality, i.e. unifying the level by which data is

Table 3.6: Description of adults census database

	Attribute	Type	Domain Size	Height
1	Age	Numeric	74	4
2	Work Class	Categorical	7	2
3	Education	Categorical	16	3
4	Country	Categorical	41	2
5	Marital Status	Categorical	7	2
6	Race	Categorical	5	1
7	Gender	Categorical	2	1
8	Salary	Sensitive	2	
9	Occupation	Sensitive	14	

generalized to achieve the privacy constraint of the compared techniques.

We start by considering a published table that satisfies 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $e = 2$. Quasi-identifiers are chosen to be Age and WorkClass where QID $= (1, 2)$. From the results shown in Fig. 3.4(a), an observed instance has a considerably high entropy loss at $[C_7]$. This clearly identifies a major privacy loss in the published table for users in this class Age $= [75, 100]$, WorkClass $= Gov$. To further understand the reason behind this loss, we refer back to the distribution of the sensitive attribute at this specific class before and after publishing. Fig. 3.4(b) shows the original versus the published distribution of the sensitive attribute. It is obvious that $[C_7]$ has some unrepresented attribute values. Hence, an observer can eliminate these values and thus gain an increased confidence about the sensitive attribute value of the user of interest. Specifically, an observer, knowing that a certain user of interest falls in the age-range Age $= [75, 100]$ and work class category WorkClass $= Gov$, can eliminate 8 possible attribute values from the Occupation's domain.

Based on the existing techniques explained earlier, a published table T' satisfying 0.1-closeness, 13-diversity and $k \geq 13$-anonymity at $e = 2$ is assumed to be privacy-preserving with these near optimum values of parameters for each PPDP technique. However, observing

the published table, we find that there is a noticeable privacy loss in $[C_2]$. More specifically, an observer, with just knowing that the user of interest is more than 50 years old, will have 100% confidence that this user's Occupation is not *Armed-Forces*. This privacy loss could be noticed using our privacy metrics. The distribution and entropy loss values of $[C_2]$ are relatively high, where $\mathscr{L}_{P_D}(S, \mathcal{C}) = [0.0125, 0.0477]$ and $\mathscr{L}_{P_E}(S, \mathcal{C}) = [0.0015, 0.0306]$. The increased distribution loss is due to a fully non-represented attribute value in $[C_2]$ of the published table.

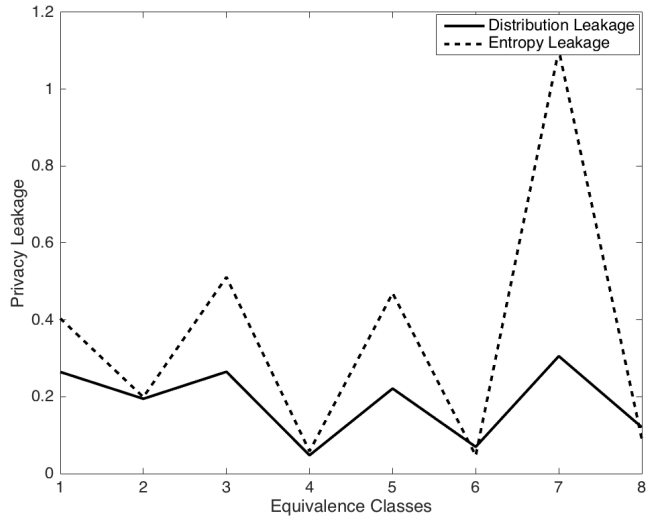
It is not necessarily an unrepresented attribute value that causes privacy loss. Fig. 3.5(b) demonstrates the original versus the published distribution for $[C_6]$ of a published table satisfying 0.5-closeness, 7-diversity and $k \geq 7$-anonymity at $e = 3$. We can see the noticeable variation in published distributions of the 8^{th} and 10^{th} attribute values $[0.0369, 0.3871]$ compared to their original distribution $[0.1077, 0.1339]$. This is expressed in our distribution loss metric shown in Fig. 3.5(a) where its value is relatively high for this specific class at 0.2853.

In addition to comparing privacy loss of different privacy levels of PPDP techniques, our work also provides a quite useful tool to compare data utility and privacy losses of different combinations of chosen quasi-identifiers in PPDP techniques. For example, let us consider four versions of a published table T' at $e = 2$. In Fig. 3.6, we compare distribution and entropy losses of the four tables while choosing a different combination of quasi-identifiers for each table, where quasi-identifiers are chosen to be $\mathsf{QID} = [(1, 2), (2, 3), (2, 4), (3, 4)]$. To satisfy the privacy conditions of the PPDP techniques, the anonymization process would decrease the number of classes in the published table and hence, the data utility decreases. In particular, Fig. 3.6 shows that anonymization process ended up with 8 classes at $\mathsf{QID} = (1, 2)$, 6 classes at $\mathsf{QID} = [(2, 3), (3, 4)]$, and 4 classes at $\mathsf{QID} = (2, 4)$. The figure illustrates the number of classes Q at each chosen combination and different levels of privacy represented in

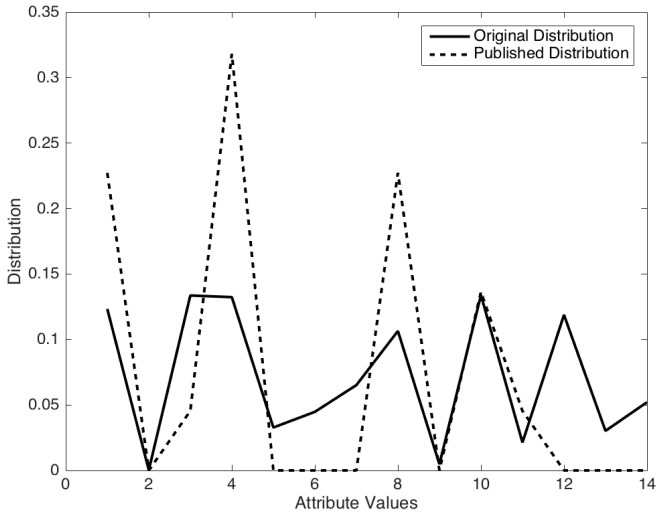
distribution and entropy losses for each class. Depending on the sensitivity of different classes formed by different combinations of quasi-identifiers, this tool gives an interesting option to adjust parameters by which a data publisher achieves the desirable privacy level with the requested data utility. Specifically, if a data publisher is more concerned about privacy of certain users that fall into $[C_3]$, then, as shown in Fig. 3.6, choosing $\mathsf{QID} = [2, 3]$ would leak too much private information about these users. Hence, according to our proposed metrics, a data-publisher can not only design the suitable data publishing technique for all users in a dataset, but also for a certain set of them.

3.5 Summary

In this chapter, we introduced two novel privacy loss quantification metrics. We elaborated the intuition behind the proposed metrics and analytically proved their correctness. Supported by insightful examples, we then showed that privacy could not be quantified based on a single metric. Based on the proposed metrics, the privacy losses of state-of-the-art PPDP techniques have been evaluated. The provided experiments demonstrate how we could gain a better judgment of existing techniques and help analyze their effectiveness in reaching privacy.

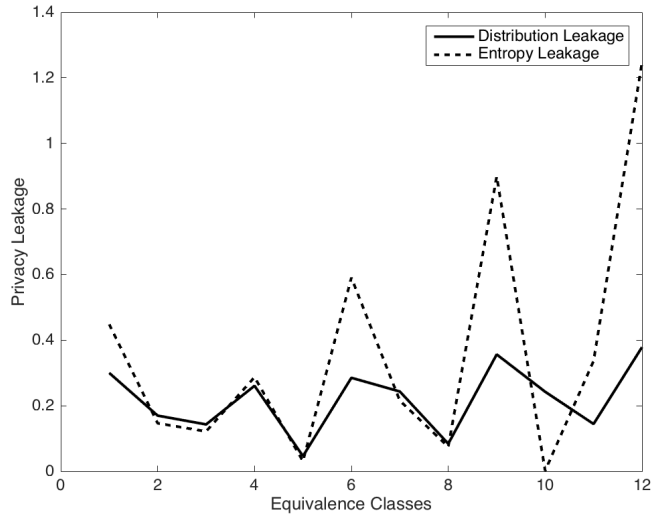


(a) Distribution and entropy losses

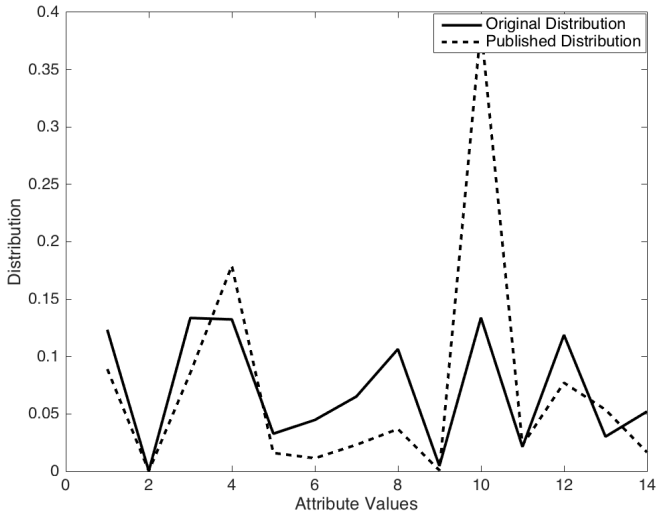


(b) Original and published distribution of a specific class

Figure 3.4: Evaluation of a table satisfying 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $e = 2$

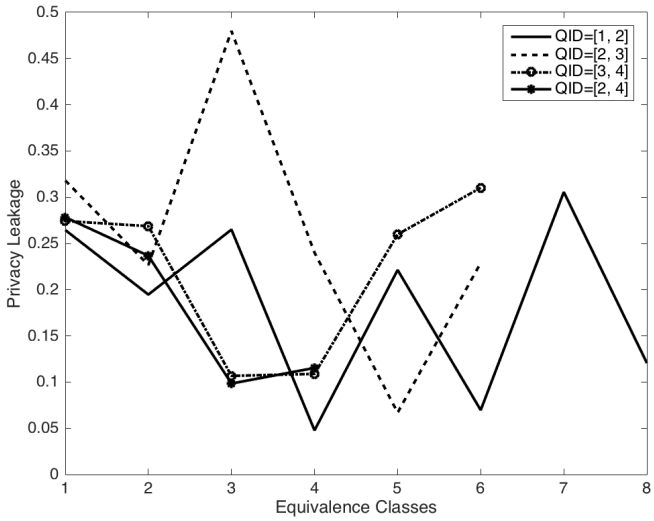


(a) Distribution and Entropy Loss

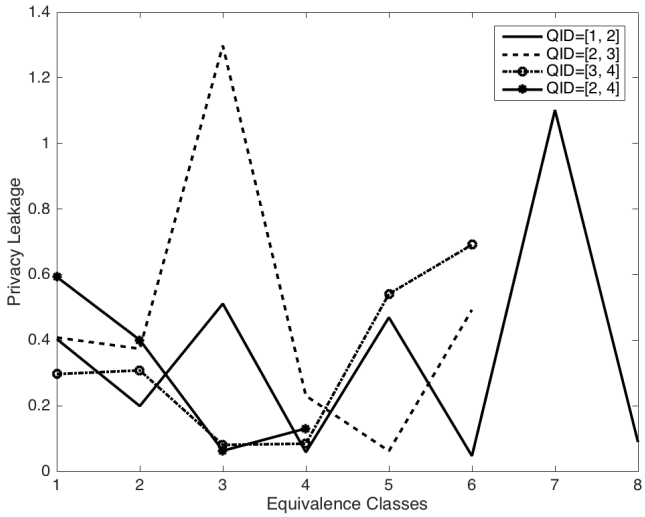


(b) Original and Published Distribution of a Specific Class

Figure 3.5: Evaluation of a table satisfying 0.5-closeness, 7-diversity, and $k \geq 7$-anonymity at $e = 3$



(a) Distribution Loss



(b) Entropy Loss

Figure 3.6: Loss at different sets of QIDs

Chapter 4

Utility-Boosting Negotiation-Based PPDD

4.1 Introduction

Privacy concerns severely limit the information provided about certain sensitive attributes. Meanwhile, DOs, sticking to privacy laws such as Health Insurance Portability and Accountability Act (HIPAA), favor individuals privacy over the public beneficiary. This results in a minimized utilization of the existing data. However, the main reason behind such miss-utilization is the lack of data disclosure techniques that provide a satisfactory tradeoff between privacy and utility of disclosed data. Data utility inevitably conflicts with data privacy. From the data utility perspective, it is best to disclose a dataset as is, while from the perspective of data privacy, it is best to disclose a mostly generalized dataset.

In some scenarios, a data recipient is in crucial need for certain attributes of interest for some decision making problems. A data recipient is usually just interested in subset of the dataset that contains the attributes of interest. These attributes will help the data recipient in making the correct decision. Consider the example of preventive prophylactic surgeries that remove an organ or gland that shows no signs of cancer in an attempt to prevent high risk individuals from developing the disease. If a patient obtains accurate information of the

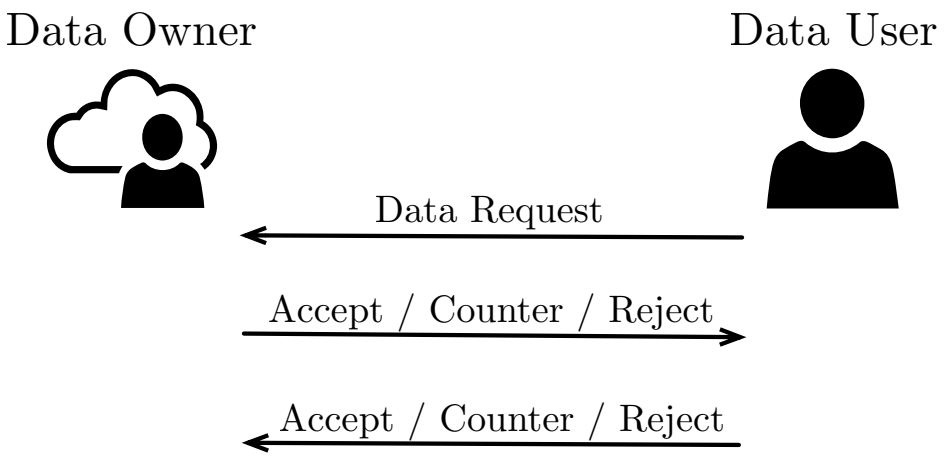


Figure 4.1: Negotiation-based data disclosure model

infection risks, they would then be able to evaluate the risks and take suitable precautions to avoid this disease.

In this chapter we propose a Utility-Boosting Privacy-Preserving Data Disclosure model (UBNB-PPDD) that redefines data utility as a function of the DU's requirements, namely, attributes of interest. The model shown in Fig. 4.1 incorporates a negotiation process between the DO and the DU in order to reach a data disclosure deal. The DU represents their requirements as utility patterns of the attributes of interest while the DO's requirements are represented as generalization policies. Based on this model we propose two UBNB-PPDD protocols that are guided by the DO's objective. The protocols provide a set of rules for the communication sessions between the DO and the DU in order to reach a data disclosure deal. The first protocol manages a negotiation process to disclose any generalized dataset that matches the DU's utility requirements and meanwhile satisfies the DO's privacy constraints. In the second protocol, the DO links the utility level of the disclosed dataset to a profit function. In the proposed model, we assume no collusion attacks. Disclosed data is protected by copyright. Data disclosure to a user is not a grant of ownership.

We propose two utility loss metrics, distribution and entropy utility loss. We exploit

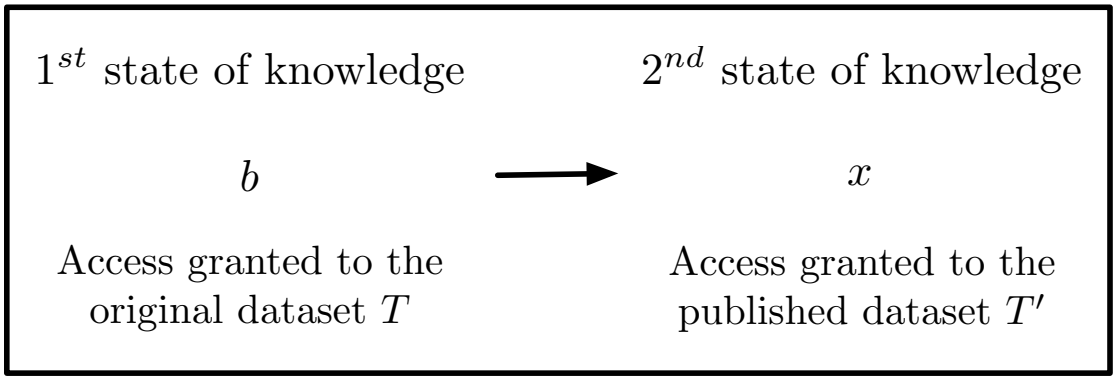


Figure 4.2: Definition of utility loss

these metrics together with the privacy loss metrics proposed in Chapter 3 to formulate the utility-privacy tradeoff and help the DO manage the decision making in data disclosure.

4.2 The Proposed Utility Quantification Metrics

It is trivial that from the data user perspective, an optimal disclosed table is a table with number of classes equal to the number of individuals in the original dataset. However, being subject to the privacy rules, the disclosed dataset loses utility in generalizing the disclosed table. Based on our model, shown in Fig. 4.2, in order to quantify the utility loss resulting from data generalization, we propose two utility metrics that build up the loss of each individual in the dataset to find the total utility loss of the disclosed table.

The first metric measures the entropy distance between the original individual distribution and the disclosed distribution of the sensitive attribute. That is, it typically measures the difference in the data user's uncertainty about the sensitive attribute value of a certain individual between two cases. In the first case, access to the original dataset is granted. In the second case, the data user only has access to the disclosed dataset which is the typical case. In other words, the utility loss metric measures how much uncertainty about sensitive

attribute value is gained after data disclosure. In this case, the original distribution b_i of an individual is a vector of zeros except for a one at the sensitive attribute value. We define the utility loss for an individual in the disclosed dataset due to generalization as follows.

Definition 11 (**Individual Entropy Utility Loss** $\mathscr{L}_{U_E}(S,[I])$)**.** The utility loss of an individual located in a certain class as a result of table generalization. This loss is given as

$$\begin{aligned}\mathscr{L}_{U_E}(S,[I]) &= \left|\sum_{i=1}^{m} b_i \log_2 \frac{1}{b_i} - \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i}\right|. \\ &= \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i}.\end{aligned}$$

As a matter of fact, if access is granted to the original dataset, the data user will have zero uncertainty about the individual's sensitive attribute value. Therefore, as shown in Definition 11, the first term of the utility loss metric is ignored. This definition can be easily extended to measure the utility loss for a class $[C_q]$. The loss of all individuals in a class $[C_q]$ is the same. Therefore, we give the following definition.

Definition 12 (**Class Entropy Utility Loss** $\mathscr{L}_{U_E}(S,[C_q])$)**.** The utility loss of all individuals falling in a certain class $[C_q]$ as a result of table generalization. This loss is given as

$$\begin{aligned}\mathscr{L}_{U_E}(S,[C_q]) &= \left|\sum_{i=1}^{m} b_i \log_2 \frac{1}{b_i} - \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i}\right|. \\ &= \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i}.\end{aligned}$$

We also simply extend this definition to measure the total utility loss of a disclosed table as a result of generalization. Assuming a disclosed table T' comprises Q equivalence

classes and the number of individuals in a class $[C_q]$ is denoted as N_q, we give the following definition.

Definition 13 (**Total Entropy Utility Loss** $\mathscr{L}_{U_E}(S, T')$)**.** The total utility loss of a disclosed table T' as a result of table generalization. This Loss is given as

$$\begin{aligned}\mathscr{L}_{U_E}(S, [T']) &= \frac{1}{N}\sum_{q=1}^{Q} N_q \left| \sum_{i=1}^{m} b_i \log_2 \frac{1}{b_i} - \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} \right|. \\ &= \frac{1}{N}\sum_{q=1}^{Q} N_q \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i}.\end{aligned}$$

We refer to the proof in our previous work in Chapter 3 to prove that the proposed utility measurement is a distance metric. Since this metric will be exploited in our model to quantify the utility loss it is also useful to have the following threshold definition as it will express each entity's constraints.

Definition 14 (γ-**Entropy Utility Loss**)**.** A disclosed table T' is said to have an γ-entropy utility loss if it has entropy utility loss $\mathscr{L}_{U_E}(S, \mathcal{C}) \leq \gamma$ for the set of all equivalence classes. That is,

$$\max(\mathscr{L}_{U_E}(S, [C_q])) \leq \gamma, \;\; q = 1, 2, \cdots, Q.$$

As we previously mentioned, a single metric is never sufficient to measure the utility loss in a multidimensional dataset. We propose another utility metric that depends on the euclidean distance between the disclosed distributions of the sensitive attribute at each class and the original conditional distributions of the sensitive attribute for each individual given these classes. The entropy metric measures the change in uncertainty of a data user about an individual's sensitive attribute value after data disclosure. The distribution metric, on the other hand, measures the deviation of the disclosed sensitive attribute values for an attribute

of interest in a certain class than the individual's attribute value.

Definition 15 (**Individual Distribution Utility Loss** $\mathscr{L}_{U_D}(S,[I])$)**.** For an individual u belonging to an equivalence class $[C_q]$, the distribution utility loss of attribute S given an equivalence class $[C_q]$ is defined as the Euclidean distance between the two distributions b and x

$$\mathscr{L}_{U_D}(S,[I]) = \sqrt{\sum_{i=1}^{m}(b_i - x_i)^2}.$$

To measure the distribution utility loss, we simply sum the utility loss of each individual in the class. The distribution utility loss in a class can then be defined as follows.

Definition 16 (**Class Distribution Utility Loss** $\mathscr{L}_{U_D}(S,[C_q])$)**.** The distribution utility loss in an equivalence class $[C_q]$ of attribute S is defined as

$$\mathscr{L}_{U_D}(S,[C_q]) = \frac{1}{N}\sum_{j=1}^{N_q}\sqrt{\sum_{i=1}^{m}\left(b_{j,i} - x_i\right)^2}.$$

Building up the distribution utility loss due to generalization, the loss for the whole disclosed dataset is the sum of losses in all classes. The total distribution utility loss in a disclosed table is defined as follows.

Definition 17 (**Total Distribution Utility Loss** $\mathscr{L}_{U_D}(S,[T']))$**.** The distribution utility loss in a disclosed table T' of attribute S is defined as

$$\mathscr{L}_{U_D}(S,[T']) = \frac{1}{N}\sum_{q=1}^{Q}\sum_{j=1}^{N_q}\sqrt{\sum_{i=1}^{m}\left(b_{j,i} - x_i\right)^2}.$$

The threshold condition for the distribution utility loss is given as follows.

Definition 18 (δ-**Distribution Utility Loss**). A disclosed table T' is said to have an δ-Distribution Utility Loss if it has distribution utility loss $\mathscr{L}_{U_D}(S, calC) \leq \delta$ for the set of all equivalence classes. That is,

$$\max(\mathscr{L}_{U_D}(S, [C_q])) \leq \delta, \;\; q = 1, 2, \cdots, Q.$$

4.3 Utility-Privacy Tradeoff Characterization

Data owners are usually opposed by the privacy concerns that result from publishing data. The data owners, therefore, have to stick to privacy constraints that protect individuals data from being revealed to adversaries. To satisfy such constraints, the published data is generalized. As the number of attributes increases this generalization optimization process becomes computationally hard. Our approach in privacy characterization and quantification not only can be used to evaluate, but can also be used to design scalable data publishing techniques.

Our goal in this section is to provide the formalization of the generalization optimization problem to achieve an optimum publishing distribution for different classes. If each of these classes achieves this publishing distribution, the privacy and utility losses are minimized for a given privacy or utility constraint. It is well understood that the generalization process might not be able to find a dataset generalization that corresponds to the recommended publishing distribution. However, we believe that providing the generalization process with such a guiding solution will substantially reduce the search scope and render the process practical.

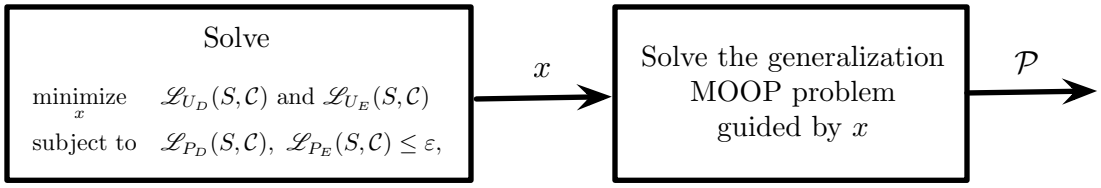


Figure 4.3: Minimum utility loss privacy-constrained data disclosure

4.3.1 Problem Formulation

So far, we have used our model to evaluate or guide the existing publishing techniques. We now investigate the more challenging task, that is formalization of the utility-privacy tradeoff in light of our proposed privacy and utility metrics. That is, we intend to use our recommended publishing distributions to solve the ultimate impracticality problem of the generalization process especially when dealing with large datasets.

Privacy and utility losses have to be considered jointly. There are three ways to approach this problem. First is the **Minimum Utility Loss Privacy-Constrained Data Disclosure**. As shown in Fig. 4.3, for a given privacy constraint ε, the goal is to find the recommended publishing distribution x that minimizes the utility loss subject to the privacy constraint.

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \mathscr{L}_{U_D}(S,[C]) \text{ and } \mathscr{L}_{U_E}(S,[C]) \\ \text{subject to} \quad & \mathscr{L}_{P_D}(S,[C_j]),\ \mathscr{L}_{P_E}(S,[C_j]) \leq \varepsilon, \\ & \sum_{i=1}^{m} x_i = 1. \end{aligned}$$

Second is the **Minimum Privacy Loss Utility-Constrained Data Disclosure**. As shown in Fig. 4.4, for a given utility constraint δ, the goal is to find the recommended publishing distribution x that minimizes the privacy loss subject to the utility constraint.

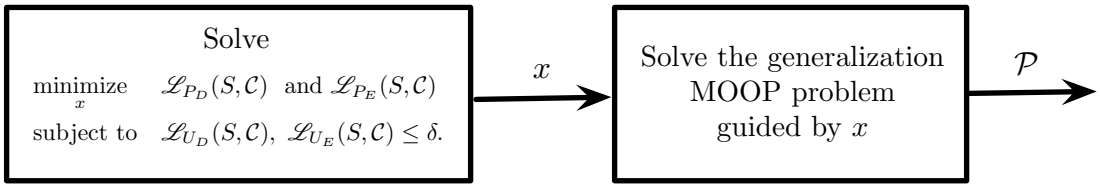


Figure 4.4: Minimum privacy loss utility-constrained data disclosure

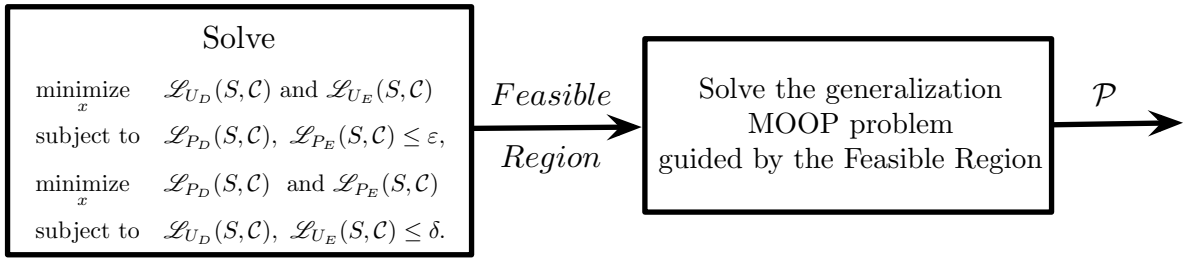


Figure 4.5: Utility-privacy loss constrained data disclosure

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \mathscr{L}_{P_D}(S,[C_j]) \ \text{ and } \mathscr{L}_{P_E}(S,[C_j]) \\ \text{subject to} \quad & \mathscr{L}_{U_D}(S,[C]),\ \mathscr{L}_{U_E}(S,[C]) \leq \delta, \\ & \sum_{i=1}^{m} x_i = 1. \end{aligned}$$

Alternatively, we may explore the feasible range of x by jointly considering the ε-privacy and δ-utility losses ranges, as illustrated in Fig. 4.6 where the feasible range is the area with green background. Thus, the third is the **Optimal Utility-Privacy Loss Constrained Data Disclosure**. As depicted in Fig. 4.5, for any given privacy and utility constraints (ε, δ), the goal is to find the **feasible region** that provides solutions for x where both privacy and utility constrains are satisfied. An optimum solution that minimizes both privacy and utility losses may then be achieved by exhaustively searching the **feasible region** space for an optimum candidate solution x.

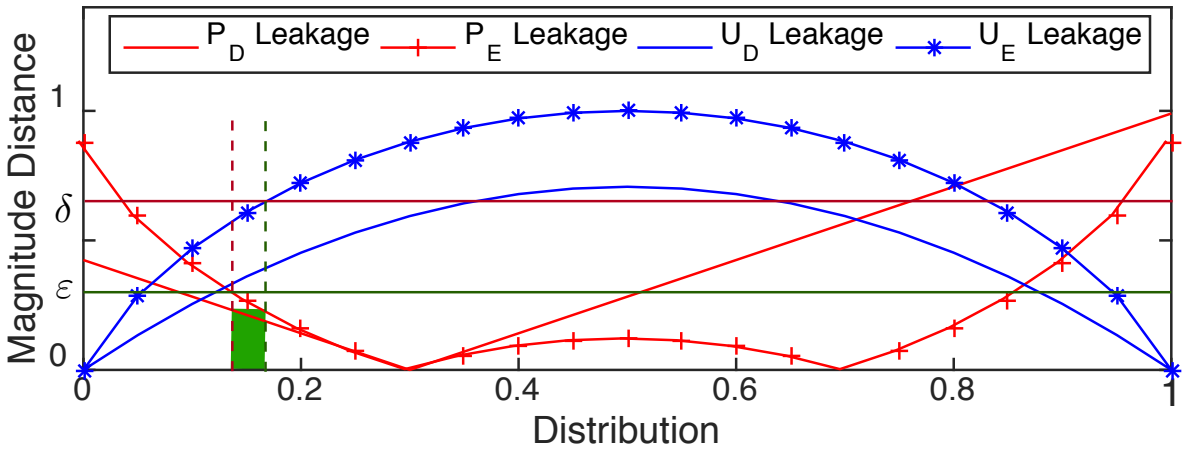


Figure 4.6: Privacy and utility tradeoffs

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \mathscr{L}_{U_D}(S,[C]) \text{ and } \mathscr{L}_{U_E}(S,[C]) \\ \text{subject to} \quad & \mathscr{L}_{P_D}(S,[C_j]),\ \mathscr{L}_{P_E}(S,[C_j]) \leq \varepsilon, \\ & \sum_{i=1}^{m} x_i = 1, \end{aligned}$$

and

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \mathscr{L}_{P_D}(S,[C_j]) \text{ and } \mathscr{L}_{P_E}(S,[C_j]) \\ \text{subject to} \quad & \mathscr{L}_{U_D}(S,[C]),\ \mathscr{L}_{U_E}(S,[C]) \leq \delta, \\ & \sum_{i=1}^{m} x_i = 1. \end{aligned}$$

These three cases are vector optimization problems. Solving these problems reveals the disclosed data distribution x that if achieved by the generalization algorithm, will lead to an optimal solution to our problem.

Losses can be computed on either an equivalence class or total losses basis to be then compared to the thresholds (ε, α) and (γ, δ). In the case of class loss basis, the thresholds are applied to each class where, the loss in each class should not exceed the threshold. In this case, the threshold is defined as the maximum accepted loss for each disclosed class. In

the case of total loss basis, the loss in the whole table should not exceed threshold. In this case, the threshold is defined as the maximum accepted loss for the disclosed table.

4.3.2 The Data Generalization Management

In order to preserve privacy, the subset dataset is then generalized according to a Multi-Objective Optimization Problem (MOOP). The generalization process takes as input the recommended publishing data distribution x and tries to find a generalization with a corresponding publishing distribution. Consequently, there exist two possible scenarios. In the first scenario, a generalization is found that corresponds to the recommended x. This generalization can be considered an optimum solution according to our privacy model. In the second scenario, the objective of the generalization process would be to find the generalization that will result in the nearest publishing distribution to the recommended x.

On one hand, this can guarantee an optimum solution in case the generalization process was able to output the recommended published distribution or a sub-optimal solution in case the output is close to the recommendations. On the other hand, these recommendations substantially reduce the amount of time and computations needed by the generalization process in case of an exhaustive search.

4.4 The Proposed UBNB-PPDD Model

The proposed negotiation-based data disclosure model manages a negotiation process where a set of communication sessions is held between the DO and the DU in order to set a data disclosure deal. More specifically, it redefines the data utility and adjusts the disclosure rules accordingly. The ultimate goal of the proposed model is to provide the DU with the

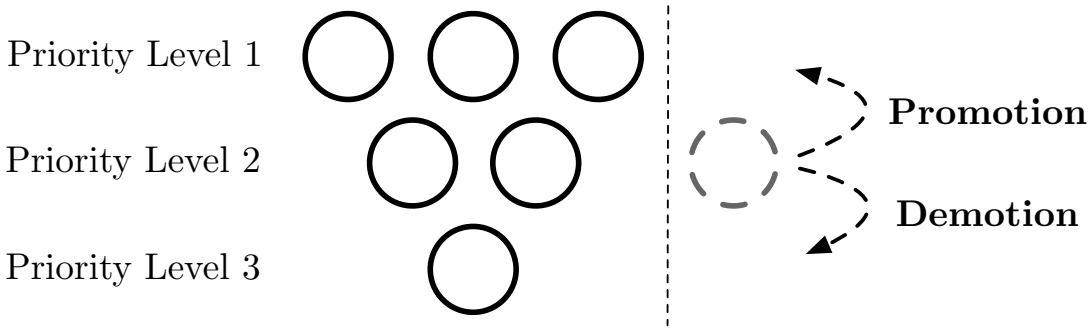


Figure 4.7: Utility pattern

expected data utility while satisfying the privacy rules of the DO. Therefore, the proposed model simply reformulates the tasks of entities in order to express their needs.

The DU might give more priority to certain attributes than others. In order to boost the data utility, the DU seeks not just the highest possible data utility in general, but rather the highest possible data utility for the attributes of interest. To express these needs, the data utility is manifested as the DU's attributes of interest divided into levels with different priorities. These priorities can be represented in a requirements priority diagram namely, the utility pattern.

Definition 19 (**Utility Pattern $\mathcal{U}$**)**.** A layered hierarchy that ranks the DU's attributes of interest. The attributes with highest priority level are located at the top layers while the attributes with the least priority level are located at the bottom.

The utility pattern can be modified by either *promotion* or *demotion* of different attributes according to their priority. As shown in Fig. 4.7, the *promotion* operation is done by the DU during the negotiation process in order to upgrade the priority level of an attribute of interest. On the other hand, the *demotion* operation is done by the DU in order to degrade the priority level of an attribute of interest.

For example, as shown in Fig. 4.8, if the DU is a medical researcher that wants to

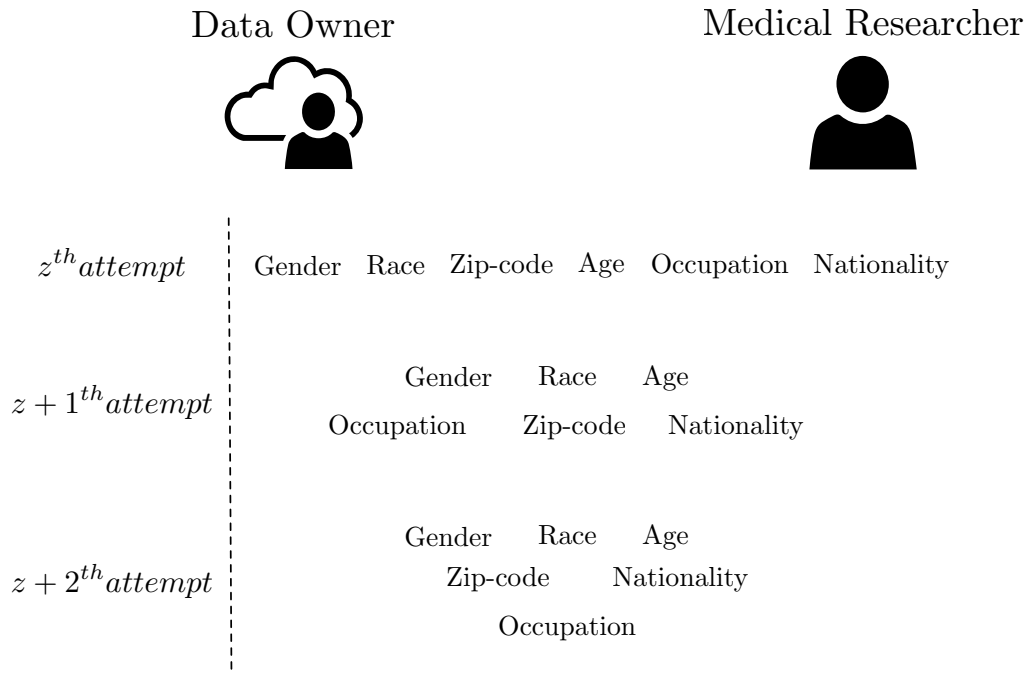


Figure 4.8: Example showing the negotiation using the utility pattern $\mathcal{U}$

evaluate the effect of some demographics such as Age, Gender, Zip-code, Race, Nationality and Occupation on a disease such as Breast Cancer. In the first negotiation attempt, the DU will have all the attributes of interest in the first priority level. If disclosing the table with these attributes of interest, as is, satisfies the DO's privacy constraints, the deal is set. Otherwise, the DU will have to rearrange the priority level orders in the second negotiation attempt by the demotion of some attributes. For instance, in our example, the researcher might have an essential need to more specific data in terms of Age, Gender and Race as compared to other attributes. Therefore, other attributes can be demoted to the second priority level in the second negotiation attempt.

In response to a DU's requested utility pattern $\mathcal{U}$, the DO recommends a generalization policy that matches $\mathcal{U}$ in compliance with the privacy constraint ε. For all possible generalizations $\mathcal{G}$ satisfying ε, we define this generalization policy as follows.

Definition 20 (**Generalization Policy** $\mathcal{P}$)**.** A policy $\mathcal{P} \in \mathcal{G}$ proposed by the DO as a re-

sponse to the utility pattern $\mathcal{U}$ requested by the DU. This policy comprises the generalization values of each attribute of interest in $\mathcal{U}$.

Consider a utility pattern $\mathcal{U}$ with d attributes of interest $\{v_1, v_2, \cdots, v_d\}$. The generalization policy is a mapping function $f: v \rightarrow v'$ that maps any attribute v to a generalized attribute v'. For example the Age attribute in Table 3.1 consists of 12 values corresponding to the ages of 12 individuals. These values are mapped to 3 classes in Table 3.2. Similarly, for all the attributes of interest, the generalization policy defines the mapping of each value from the original to the disclosed table.

The generalization policy recommended by the DO will not essentially satisfy the DU's expected data utility. Therefore, using our proposed utility pattern, generalization policy, utility and privacy loss metrics, both the DO and the DU go through a negotiation process to set the data disclosure deal. Details of this negotiation process are elaborated in Section 4.5.

Throughout the rest of this section, for the reader's convenience, we will merge the two privacy loss metrics terms $\mathscr{L}_{P_D}(S, \mathcal{C})$ and $\mathscr{L}_{P_E}(S, \mathcal{C})$ into $\mathscr{L}_P(S, \mathcal{C})$, and the two utility loss metrics terms $\mathscr{L}_{U_D}(S, \mathcal{C})$ and $\mathscr{L}_{U_E}(S, \mathcal{C})$ into $\mathscr{L}_U(S, \mathcal{C})$. We will also use ε instead of (ε, α) to express the privacy thresholds, and γ instead of (γ, δ) to express the utility thresholds.

To decide whether an offer can be accepted or not, the DO computes the privacy and utility losses in order to check if the response $\mathcal{P}$ to a DU's offer satisfies the requirements and constraints. Losses can be computed on either an equivalence class or total losses basis to then be compared to the thresholds ε and γ. In the case of class loss basis, the thresholds are applied to each class where the loss in each class should not exceed the threshold. In the case of total loss basis, the loss in the whole table should not exceed the threshold.

Suppose that for a given privacy threshold ε, there exist g possible generalizations $\mathcal{G}$

that satisfy the privacy constraint. The ultimate goal is to find, out of these generalizations, the one that satisfies the DO's objective. After satisfying this objective, the DO now has a recommended generalization policy that is guided by the DU's utility pattern and meanwhile satisfies the objective.

For a given ε, the DO has two possible objectives. The first is to disclose the data with the best (or any) utility corresponding to the least (or any) possible generalization within the privacy threshold. The second is to control the utility loss of the disclosed data for the purpose of linking it to a profit function.

The first proposed scenario is disclosing the dataset that satisfies the privacy constraints disregarding how much utility loss it provides as long as it satisfies the DU's requested utility pattern $\mathcal{U}$. We name this the *Flat Rate Objective*. Another proposed scenario is linking the data utility loss to a profit function. That is, for a requested DU's utility pattern $\mathcal{U}$ and privacy constraint ε, there exist g possible generalizations $\mathcal{G}$ that satisfy the privacy constraint. Each of these generalizations provides a different level of data utility loss $\mathscr{L}_U(S, [C])$. As shown in Fig.4.9, the profit decreases with the increase in the data utility loss. We name this the *Variable Rate Objective*. The profit is hence a function of the data utility loss,

$$\mathsf{Profit} = f(\mathscr{L}_U(S, [C])).$$

The profit function is typically determined by the DO depending on the data value in the market. These two objectives will drive the design of our two UBNB-PPDD protocol versions proposed in Section 4.5.

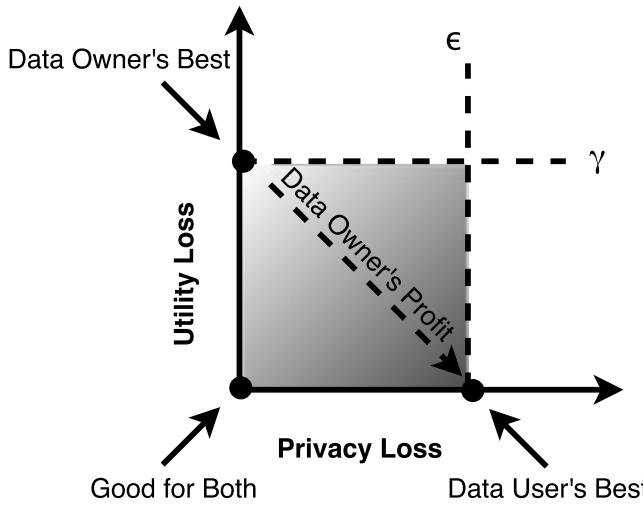


Figure 4.9: Diagram showing utility-privacy tradeoff and the DO's profit

4.5 The Proposed UBNB-PPDD Protocol

In this section we describe two variants of the proposed UBNB-PPDD protocol. Based on the proposed model, the protocol relies on a negotiation process between the DO and the DU guided by both entities' needs and expectations. These rules control the design of the previously mentioned tools, namely, the utility pattern $\mathcal{U}$ and generalization policy $\mathcal{P}$.

Let $z = 1, \cdots, Z$ represent the z^{th} negotiation session between the DO and the DU where Z is the maximum number of negotiation sessions. Throughout the negotiation process, DU modifies the requested utility pattern $\mathcal{U}_z$ by either promoting or demoting different required attributes according to their priority level. Also the DO modifies the generalization policy $\mathcal{P}_z$ through modifying the mapping function by either increasing or decreasing the range by which each attribute is generalized. This continues until both entities set on a data utility level that matches the DU requirements on one hand and a data privacy level that satisfies the DO constraints on the other hand. We note that both entities can terminate the negotiation at any time.

We introduce the two variants of the UBNB-PPDD protocol shown in Fig. 4.10. In

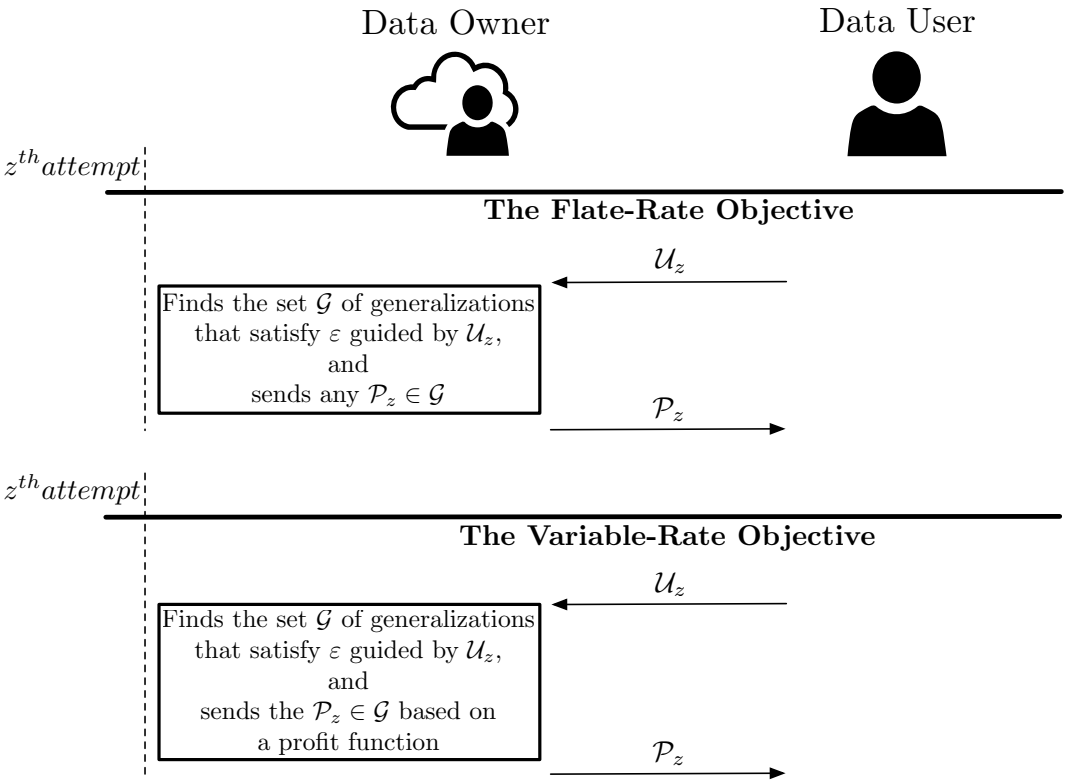


Figure 4.10: The UBNB-PPDD protocols

the first version, we employ the flat rate DO's objective while in the second we employ the variable rate DO's objective.

4.5.1 The Flat Rate UBNB-PPDD

The DU submits an *offer* by sending a requested data utility pattern $\mathcal{U}_z$. If the DO accepts the requested $\mathcal{U}_z$ as is, the DO sends *any* generalization policy $\mathcal{P}_z$ that satisfies the privacy constraint $\mathscr{L}_P(S, \mathcal{C}) \leq \varepsilon$. The DU then reviews $\mathcal{P}_z$ and either responds with an *offer accept* to make a deal or an *offer reject* to refuse it and sends a modified utility pattern if interested. If no $\mathcal{P}_z$ that matches the ε is found for the requested $\mathcal{U}_z$, the DO refuses $\mathcal{U}_z$. The DU either modifies the utility pattern and presents a new offer or does a *negotiation termination* due to the failure of reaching a suitable deal. The flat rate UBNB-PPDD protocol is summarized in Algorithm 1.

Algorithm 1 Flat rate UBNB-PPDD

1. DU sends a request $\mathcal{U}_z$ to the DO.
2. DO responds with any $\mathcal{P}_z \in \mathcal{G}$ that matches the requested $\mathcal{U}_z$ and satisfies $\mathscr{L}_P(S, \mathcal{C}) \leq \varepsilon$.
3. If satisfied by $\mathcal{P}_z$, DU responds with an *offer accept*.
4. Otherwise, DO responds with another $\mathcal{P}_{z+1} \in \mathcal{G}$ as a counter offer.
5. If no $\mathcal{P}_{z+1}$ exists, DO responds with an *offer reject* or a *negotiation termination*.
6. If the response is an *offer reject*, if interested, the DU sends a new relaxed request $\mathcal{U}_{z+1}$. Otherwise, DU does a *negotiation termination*.
7. Repeat steps 2, 3, 4, 5, and 6.

4.5.2 The Variable Rate UBNB-PPDD

In the variable rate UBNB-PPDD scenario, the DU submits an *offer* by sending a requested $\mathcal{U}_z$. The DO finds the set $\mathcal{G}$ of the generalization policies that matches the requested $\mathcal{U}_z$ and meanwhile satisfies $\mathscr{L}_P(S, \mathcal{C}) \leq \varepsilon$. The DO can either accept the requested $\mathcal{U}_z$ as is, or refuse it. In the first case, if the request is accepted, the DO computes the utility loss $\mathscr{L}_U(S, \mathcal{C})$ and the $Profit$. The DO then sends an *optimized* $\mathcal{P}_z$ that satisfies the privacy constraint $\mathscr{L}_P(S, \mathcal{C}) \leq \varepsilon$ and matches the expected profit. The DU then reviews $\mathcal{P}_z$ and either responds with *offer accept* to make a deal or *offer reject* to refuse it and sends a modified utility pattern if interested. In the second case, if the DO refuses due to the non-existence of a generalization policy that matches the privacy constraint for the requested $\mathcal{U}_z$, the DU either modifies the utility pattern and presents a new offer or does a *negotiation termination*. The variable rate UBNB-PPDD protocol is summarized in Algorithm 2.

We note that a flat rate objective saves computations at the DO's side where the DO is not required to optimize the generalization process to be linked to the profit function. However, this does not have any guarantees about neither achieving the best possible data

Algorithm 2 Variable rate UBNB-PPDD

1. DU sends a request $\mathcal{U}_z$ to the DO.
2. DO finds the set $\mathcal{G}$, computes $\mathscr{L}_U(S, \mathcal{C})$ and the Profit, and responds with an optimized $\mathcal{P}_z$.
3. DU reviews $\mathcal{P}_z$ and either responds with an *offer accept* or *offer reject* and sends a modified $\mathcal{U}_{z+1}$ if interested.
4. If no $\mathcal{P}_z \in \mathcal{G}$ exists, DO responds with an *offer reject* or a *negotiation termination*.
5. If the response is an *offer reject*, if interested, the DU sends a new relaxed $\mathcal{U}_{z+1}$. Otherwise, DU does a *negotiation termination*.
6. Repeat steps 2, 3, 4 and 5.

utility for the DU nor the best possible profit for the DO.

4.6 Empirical Analysis and Simulation Results

This section is divided into two parts. In the first part, based on our findings, we introduce an empirical example to help understand the implications of the proposed utility metrics. In the second part, through simulations on the US census dataset from the UC Irvine machine learning repository [66], we show how the proposed metrics can enable the DO to evaluate the privacy and utility of a disclosed dataset.

4.6.1 Empirical Analysis

In this subsection we introduce an empirical example to help understand the capabilities of the utility and privacy metrics in evaluating the utility and privacy losses of disclosed classes.

Example 12. In the example from [4], the original impatient dataset is given in Table 4.1 and the 4-anonymous impatient dataset is given in Table 4.2. For these two tables, the probability

Table 4.1: Original dataset

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Virus Infection
4	13053	23	American	Virus Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Virus Infection
8	14850	49	American	Virus Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Table 4.2: 4-anonymous impatient micro-data

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Virus Infection
4	130**	<30	*	Virus Infection
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Heart Disease
7	1485*	≥40	*	Virus Infection
8	1485*	≥40	*	Virus Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

distribution for the three diseases is $\left(\frac{3}{12}, \frac{4}{12}, \frac{5}{12}\right)$. In this case, the privacy loss $\mathscr{L}_P(S, \mathcal{C})$ for individuals within the first, second and third equivalence classes is $[0.5137, 0.2357, 0.7619]$, while the utility loss $\mathscr{L}_U(S, \mathcal{C})$ is $[0.707, 0.77, 0]$ respectively.

Our findings for Table 4.2 reveal that patients under 30 have *Heart-Disease* or *Virus-Infection* with equal probability. The scheme provides $\mathscr{L}_P(S, [C_1]) = 0.51$ and $\mathscr{L}_U(S, [C_1]) = 0.707$. For patients over 40, $\frac{1}{4}$ have *Cancer*, $\frac{1}{4}$ have *Heart-Disease* and $\frac{1}{2}$ have *Virus-Infection*. The scheme provides $\mathscr{L}_P(S, [C_2]) = 0.23$ and $\mathscr{L}_U(S, [C_2]) = 0.77$. Finally, patients in their 30s, all have *Cancer*. The scheme returns $\mathscr{L}_P(S, [C_3]) = 0.76$ and $\mathscr{L}_U(S, [C_3]) = 0$.

4.6.2 Simulation Results

Simulation results give us an insightful understanding of utility and privacy losses and how the negotiation can be handled in our proposed UBNB-PPDD protocol. Specifically, the DO is able to analyze the utility and privacy losses for different combinations of QIDs and interpret the losses in each class to determine which classes leak more privacy or provide more utility. Simulations are done on a sample of the US census dataset. After eliminating records with missing values, we have a total of 30,162 records. Following the work in [4], as shown in Table 3.6, we utilize only 9 attributes, 7 of which form the set of possible quasi-identifiers while Occupation and Salary form the set of possible sensitive attributes. We adopt the Incognito algorithm [41] for generating the generalized tables with Occupation as the sensitive attribute. The number of quasi-identifiers QIDs is represented by the variable e that takes values from 1 to 7 with the same order in Table 3.6.

We start by considering a disclosed table satisfying 0.5-closeness, 6-diversity, and 6-anonymity at $e = 2$. Quasi-identifiers are chosen to be Age and WorkClass where QID $= (1, 2)$. From the results shown in Fig. 4.11(a), an observed instance has a considerably high privacy

loss for individuals in $[C_7]$. To further understand the reason behind this loss, we refer to Fig. 4.11(b) showing the distribution of the Occupation in the original table versus the distribution at this specific class after disclosure. It is obvious that $[C_7]$ has some unrepresented attribute values. Hence, an observer can eliminate these values and thus gain an increased confidence about the Occupation of the individual of interest. Specifically, an observer, knowing that a certain individual of interest falls in the age-range $\mathsf{Age} = [75, 100]$ and work class category $\mathsf{WorkClass} = Gov$, can eliminate 8 possible attribute values from the Occupation domain. Also as shown in Fig. 4.11(a), this class has the least utility loss. This is also justifiable by the fact that the DU gains a high level of certainty about the sensitive attribute values of individuals in this class where only 6 out of 14 sensitive attribute values are represented. Thus, the DO can use the privacy and utility metrics to manage the negotiation process. In particular, the DO is able to control the privacy and utility loss levels of different classes depending on the threshold values and the expected profit.

For the requested DU's utility patterns $\mathcal{U}$s, the DO can analyze the utility and privacy losses of any dataset generalization before disclosure. For example, in Fig. 4.12, we compare privacy and utility losses of four disclosed tables at $e = 2$ while choosing a different combination of quasi-identifiers for each table. Quasi-identifiers are chosen to be $\mathsf{QID} = [(1, 2), (2, 3), (2, 4), (3, 4)]$.

To satisfy the privacy constraints, for different requested utility patterns with different attributes of interest, the generalization would decrease the number of classes in the disclosed table and hence, the data utility decreases. Fig. 4.12 also illustrates the number of classes Q at each chosen combination of QIDs and different levels of privacy and utility loss for each class. The generalization ended up with 8 classes at $\mathsf{QID} = (1, 2)$, 6 classes at $\mathsf{QID} = [(2, 3), (3, 4)]$, and 4 classes at $\mathsf{QID} = (2, 4)$. Depending on the sensitivity of different classes

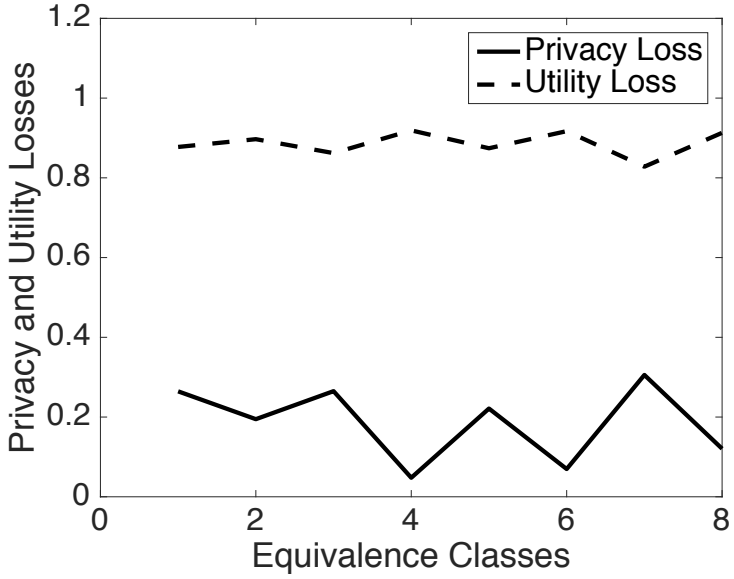
formed by different combinations of QIDs, the DO can select the generalization that achieves the desirable privacy level for a requested $\mathcal{U}$. For instance, if a DO is more concerned about privacy of certain users that fall into $[C_3]$, then choosing QID $= [2, 3]$ would leak considerable private information. Also for these specific QIDs the class $[C_3]$ has a low utility loss level. Hence, according to the metrics, a DO can not only design the suitable data disclosure technique for all individuals in a dataset, but also for a subset of them.

Let us also consider 3 versions of a disclosed table T', with different privacy loss levels, at $e = 3$, 5 and 7. As shown in Fig. 4.13, QIDs are chosen to be QID $= [(1, 2, 3), (1, 2, 3, 4, 6), (1, 2, 3, 4, 6, 7, 8)]$. This is useful where we can see the tradeoff between the utility and privacy and how different utility patterns can affect the disclosed data utility levels for a given privacy constraint.

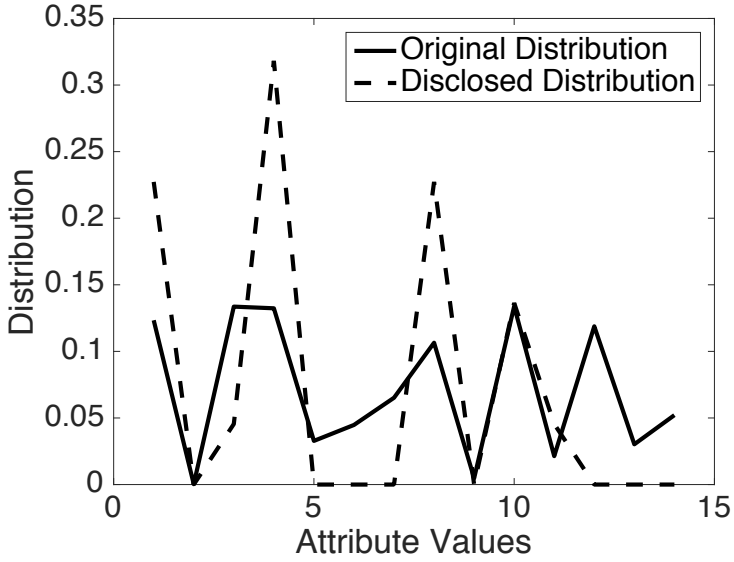
4.7 Summary

In this chapter, we introduced two data utility loss metrics. Using these metrics and the previously proposed privacy loss metrics we were able to practically address the utility-privacy tradeoff problem. An optimization problem was formulated to find a set of recommended publishing class distributions that provide an optimally minimized data utility loss for a given privacy constraint or vice versa. The generalization process of the published dataset can then be guided by the recommended class distributions. A utility-boosting privacy-preserving data disclosure model that redefines the data utility based on the DU's perspective is then proposed. Based on this model we incorporate our utility and privacy metrics to propose two versions of a privacy-preserving data disclosure protocol. The protocol sets rules for the negotiation between the DO and the DU in order to set a data disclosure deal.

The proposed protocol inherently boosts the data utility from the DU's perspective with the satisfaction of the DO's privacy constraint and monetary objectives.



(a) Privacy and Utility Losses



(b) Original distribution vs. disclosed distribution of class $[C_7]$

Figure 4.11: Evaluating privacy and utility losses of a table satisfying 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $e = 2$

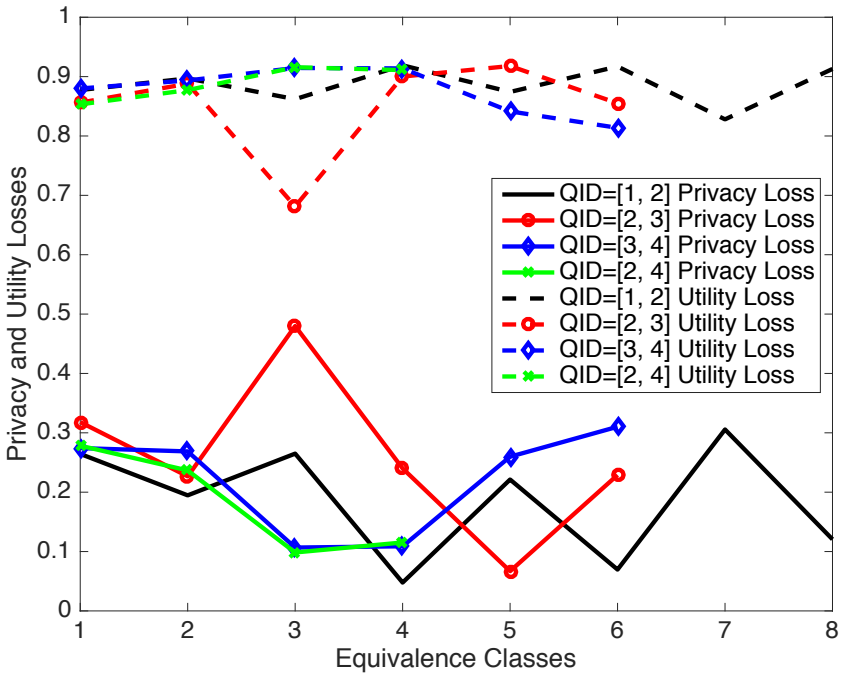


Figure 4.12: Privacy and utility losses at different sets of QIDs

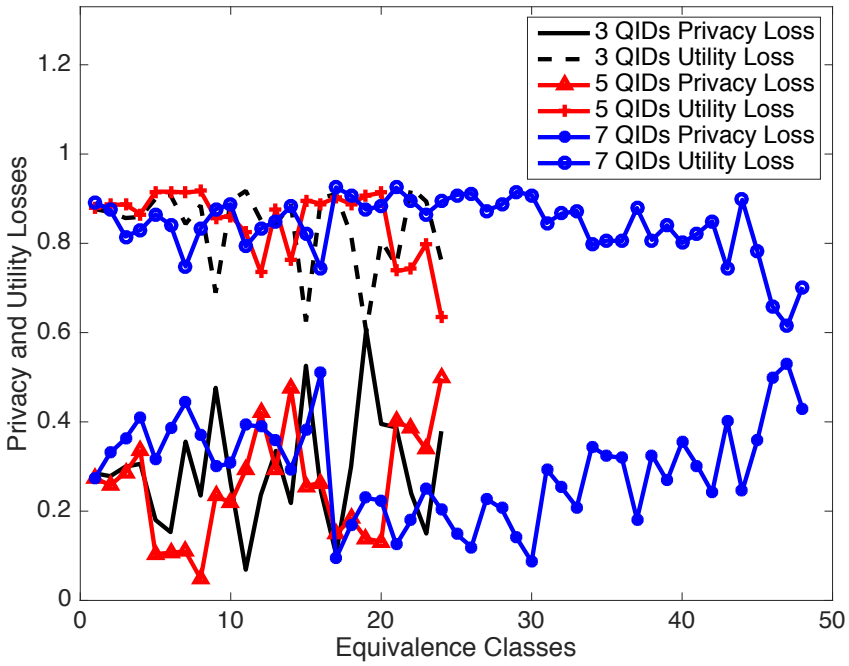


Figure 4.13: Utility-privacy tradeoff for different attributes of interest at $e = 3, 5$ and 7

Chapter 5

Privacy Preserving Data Publishing for Machine Learning Applications

5.1 Introduction

There are many examples, where sharing a public dataset could benefit the publisher as well as the broader community, including sharing security incidents, genetic information, demographic data [67–69]. However, including individuals' data in published datasets comes with its own risks and concerns. After a series of privacy loss incidents that took place over the previous years, the research community started to shed some light on this problem and uncover the hidden risks of publishing datasets without having a clear understanding of the possible unprecedented privacy losses. Not long after the researchers recognized the importance of data privacy, did the governments and international community start to put restrictions and pass laws to manage data publishing.

In order to protect data privacy from different attacks, many privacy-preserving techniques and strategies have been proposed to meet different individual's privacy and utility requirements. In PPDP, once the data is published, it is available for any type of analysis. The concern then becomes the sensitive attribute disclosure of individuals participating in the dataset. All syntactic privacy-preserving data publishing techniques typically aim

at protecting individual's privacy against sensitive attribute disclosure. This form of privacy breach is different and incomparable to learning whether an individual is included in the dataset, which is the focus of differential privacy [36]. k-anonymity, ℓ-diversity, and t-closeness, as well as our proposed privacy framework, are examples of syntactic privacy models that attempt to protect individual's privacy while minimizing impact on published data utility.

The fear of exposing private user data has constrained the ability to publish informative datasets and share them with other entities. Machine learning, for example, a technology that relies solely on data, will intuitively be negatively affected by such privacy restrictions. In ML, existing data is used to train different models which are then used for purposes of prediction, regression, or clustering of new data. The more informative the existing data is, the more accurate the trained model becomes at performing the specified task on new data. Thus, while certain levels of data privacy can be achieved using different approaches such as anonymization, perturbation, or anatomization, no matter what approach is used, it is generally believed to directly influence the accuracy of the trained ML models. It is unfortunate that this intuition, to the best of our knowledge, has never been translated into a solid framework with quantifiable measures. This is a scheme that, if exists, would help both data owners and users, that are willing to incorporate data in ML systems, to understand the value of the data after being generalized to satisfy privacy constraints. That is, enabling the evaluation of published data utility in ML systems and formally expressing the utility-privacy tradeoff.

A typical dataset would comprise 3 kinds of attributes; personally identifying, quasi-identifiers, and sensitive attributes. Personally identifying attributes are typically removed from datasets. To satisfy the privacy constraints, quasi-identifiers are generalized to avoid

de-identification of individuals in datasets. The generalization will naturally lead to dividing the dataset into classes. According to our characterization of data privacy in [70], maximizing data privacy aims at minimizing the distribution and entropy distances between the distribution of the sensitive attributes in the original dataset and its distribution in each published class. Maximizing data utility, on the other hand, aims at maximizing the accuracy of the ML model trained on generalized data as to, optimally, reach the accuracy of a model that is trained on original data. Our ultimate goal is to provide the metrics by which a data owner can quantify both data privacy and utility losses and can then manage the utility-privacy tradeoff in ML applications.

Our work answers some essential questions such as; what is the effect of data privacy preservation on results of machine learning models?, how to quantify data utility loss from a machine learning perspective?, how can the data utility-privacy tradeoff be quantitatively expressed using solid metrics?, and, most interestingly, if two datasets provide almost the same accuracy, why publish the more privacy leaking dataset?

In this chapter, we provide a framework to manage privacy preserving data publishing for machine learning applications. We also propose a quantifiable approach to measure the privacy and utility losses from a ML perspective. As shown in Figure 5.1, the approach incorporates an iterative algorithm that trains ML models on different datasets that satisfy different levels of data privacy constraints. A classification accuracy based data utility metric is also formulated to measure the drop in data utility as a result of obeying the privacy constraints. The framework is finally tested on an employees' Office365 login dataset, from Barracuda Networks. The dataset is used to train a proposed privacy preserving ATO detector where a ML model is designed to classify the fraudulence of login attempts. The results express variations in models' accuracy in binary classification of logins when trained on dif-

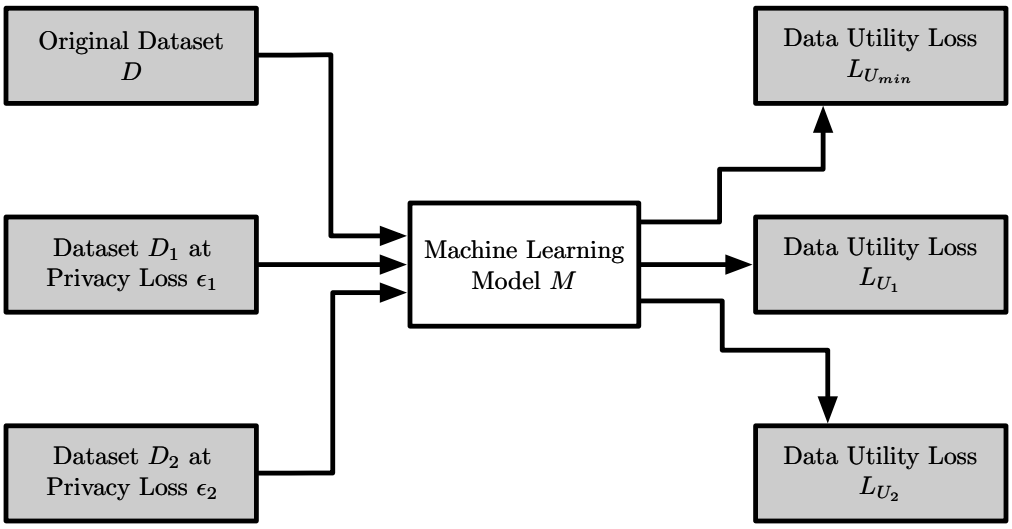


Figure 5.1: The proposed system model

ferent datasets that satisfy different privacy constraints. The proposed privacy framework enables a data publisher to quantitatively manage the utility-privacy tradeoff and provide answers to some interesting research questions.

The rest of this chapter is arranged as follows. In Section 5.2, we conduct a qualitative review of the literature. Section 5.3 presents some preliminaries. In Section 5.4, we introduce our privacy loss interpretation and the utilized privacy metrics. A brief explanation of machine learning models and their evaluation is introduced in Section 5.5. The proposed model to compute data utility loss for ML applications is presented in Section 5.6. In Section 5.7, we introduce our privacy preserving ATO detector and demonstrate the results of applying the proposed data publishing framework on real industry data. We finally draw our conclusion and suggest some open problems for future work in Section 5.8.

5.2 Related Work

Machine learning is an application of Artificial Intelligence (AI) that uses experience to make accurate predictions. This experience is represented in the form of data that a ML

model learns from. While this data might, in some cases, be non-sensitive, it is otherwise in most cases. This is generally the case with ML applications that include individuals' data. In order to shield individual privacy in the context of big data, different anonymization techniques have conventionally been used. The most relevant techniques are k-anonymity, ℓ-diversity, and t-closeness [2–6]. While k-anonymity protects identity disclosure of individuals by linking attacks, it is insufficient to prevent attribute disclosure with side information. By combining the released data with side information, it makes it possible to infer the possible sensitive attributes corresponding to an individual. To deal with this issue, ℓ-diversity was introduced in [4]. However, as stated in [5], ℓ-diversity limits the adversarial knowledge, while it is possible to acquire knowledge of a sensitive attribute from generally available global distribution of the attribute. In [5], the Earth Mover Distance (EMD) [12] is used as a metric to compute privacy loss, which is represented as the information gain for a specific individual over the entire population. However, the value t is an abstract distance between two distributions that does not have an intuitive relation with privacy loss. Moreover, we believe that the privacy loss cannot be just quantified by a single metric. In [70], authors provide a tuple of privacy loss metrics that tackle the limitations of all the mentioned metrics. Cryptographic approaches to achieve privacy preserving machine learning have been widely investigated in literature [71–75]. However, they mostly fail to provide practical and scalable solutions.

As a result of applying privacy constraints, the published data loses some of its value. Thus, it becomes essential to study the data utility loss resulting from these constraints and attempt to find quantifiable frameworks that can, reliably, capture these losses. Consequently, the tradeoff between utility and privacy has recently grabbed the research community's attention. Multiple approaches have been proposed to model this tradeoff [76,

77]. Information-theoretic frameworks that promise an analytical model guaranteeing tight bounds of how much utility is possible for a given level of privacy and vice-versa are presented in [78] and [25]. In [79], a game theoretic approach is presented where the authors model the interactions among data owners and users as a game, and propose a general approach to find the Nash equilibrium of the game. An interactive algorithm for data publishing is proposed in [80], where the tradeoff is managed through negotiable offers between DOs and DUs.

Although, the mentioned approaches provide solid frameworks for different use cases, none of them address the tradeoff from a machine learning perspective. While the vast majority of the published datasets are expected to be included in some sort of machine learning application, it is very appealing to understand the real value of generalized data when a ML model is conducted upon them. Moreover, finding a quantifiable set of metrics that can capture the degradation in models' performance due to privacy preservation techniques becomes crucial.

5.3 Preliminaries

Data Disclosure Model Some attributes can uniquely identify the individuals such as the social security or the driving license numbers. These attributes are referred to as *explicit-identifiers*. Some of the attributes are non-sensitive. These attributes are generally referred to as *quasi-identifiers*. Sensitive attributes may include information such as Disease and Salary. When datasets are published, all explicit-identifiers are removed.

It is worth to note that differential privacy models focus on membership privacy. As described in [81], differential privacy aims at answering queries while simultaneously ensur-

ing privacy of individual records databases. These models serve Privacy Preserving Data Mining (PPDM) which aims at performing data mining tasks on a set of private databases owned by different parties. In a typical PPDM scenario, the data owner maintains control over the data and does not publish it. Instead, the owner responds to previously known types of queries on the data, and ensures that the answers provided do not violate the privacy of the data subjects. This is typically achieved by adding noise to the data, and it is necessary to know the analysis to be performed in advance in order to adjust the level of added noise. This approach contradicts with the main objective of the PPDP techniques that ought to make the published data less precise than the original data but semantically truthful and hence preserve the integrity of the data. Differential privacy in different machine learning applications is investigated in [82–90].

Table Generalization To satisfy the privacy constraints, data publishing techniques apply some generalizations to the quasi-identifiers QIDs to avoid linking individuals to records in the table. Any value in the original table is mapped to a generalized value in the published table following a certain mapping function. After generalization, the published Table T' is divided into a set $\mathcal{C} = \{[C_q]\}_{q=1}^{Q}$ of Q equivalence classes. Let $S = \{s_i\}_{i=1}^{m}$ be the set of all m attribute values of a sensitive attribute $S \in \mathcal{A}$. The estimated initial distribution of S for equivalence class $[C_q]$ is given as $a_{S,[C_q]} = (a_1, a_2, \cdots, a_m)$. The published distribution of S in an equivalence class $[C_q]$ is given as $x_{S,[C_q]} = (x_1, x_2, \cdots, x_m)$. Throughout the rest of this chapter, we denote $a_{S,[C_q]}$ as a and $x_{S,[C_q]}$ as x.

5.4 The Privacy Loss Metrics

Following our work in [70], our approach to quantify privacy depends on quantifying the information loss between two adversary's states of knowledge. At the first state, based on public information of sensitive attribute's distribution, an adversary has some prior belief about the sensitive attribute value of an individual. This prior belief a is based on the probability distributions of attributes and joint distributions of their combinations. After publishing, an adversary moves to the second state of knowledge, the posterior belief x, that is the conditional distribution of sensitive attribute given combinations of published attributes. The amount information gained by the adversary after publishing is the privacy loss that we need to capture. We believe that matching the published distribution x to the original distribution a would indeed achieve better privacy. Therefore we give the following definition.

Definition 21 (**Distribution Privacy Loss** $\mathscr{L}_{P_D}(S, [C_q])$)**.** For an individual u_n in an equivalence class $[C_q]$, the privacy loss of attribute S given an equivalence class $[C_q]$ is defined as the Euclidean distance between the two distributions a and x,

$$\mathscr{L}_{P_D}(S, [C_q]) = \sqrt{\sum_{i=1}^{m} (a_i - x_i)^2}. \tag{5.1}$$

While the distribution privacy loss measures the overall divergence of attribute values distribution from one state to the other and thus, captures the amount by which privacy of an attribute is leaked, it does not give a sufficient implication about privacy loss of individuals carrying different attribute values. Specifically, a small distribution loss in the published table might lead to a critical decrease in the amount of uncertainty of an adversary

about the attribute value of a certain individual of interest. This motivates us to think of an information theoretic metric that would capture this change of adversarial uncertainty before and after publishing. Hence, we propose the following privacy metric.

Definition 22 (**Entropy Privacy Loss** $\mathscr{L}_{P_E}(S, [C_q])$)**.** For an individual u_n belonging to an equivalence class $[C_q]$, the entropy loss is defined as

$$\mathscr{L}_{P_E}(S, [C_q]) = \left| \sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i} - \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} \right|. \tag{5.2}$$

Note that the entropy privacy loss reaches maximum $\log_2 m$ when the original distribution is uniform and the published distribution is $x_i = 1$ for some attribute value i and 0 for other attribute values. This is easily explained as a transition in the adversarial belief, from a state where the adversarial uncertainty about the attribute value of an individual of interest in a given class is maximum, to a state where they become 100% confident about the attribute value of this individual. We note that maximum entropy of attribute values in the published dataset, which corresponds to uniform distribution, does not necessarily achieve the maximum privacy. This kind of distribution can be optimum if the background information of an adversary is ignored. However, given that an adversary has some prior belief about original attribute values distributions, it is best to maintain the same entropy level after publishing.

Upon data publishing, the DO's objective is to keep the privacy loss below a predetermined level.

Definition 23 ($(\varepsilon,\ \alpha)$-**Privacy Loss**)**.** A published table T' has an $(\varepsilon,\ \alpha)$-privacy loss if $\mathscr{L}_{P_D}(S, \mathcal{C}) \leq \varepsilon$ and $\mathscr{L}_{P_E}(S, \mathcal{C}) \leq \alpha$ for the set of all equivalence classes. That is $\max(\mathscr{L}_{P_D}(S, [C_q])) \leq \varepsilon$, and $\max(\mathscr{L}_{P_E}(S, [C_q])) \leq \alpha,\ \ q = 1, 2, \cdots, Q$.

5.5 Machine Learning Algorithms and their Evaluation

5.5.1 Training the Classifier

Learning algorithms typically rely on some existing labeled data points to train a classifier that is intended to classify some future, unlabeled, data points. As shown in Fig. 5.2, a classifier's training process starts by splitting labeled data into training and test data. Through cross-validation, training data is then split into validation and training samples. The training samples are then input to the feature extraction and selection phase. After features are selected, data feature vectors and validation data samples are fed into the classifier for training and hyper-parameter optimization, if exists, depending on the chosen machine learning classification algorithm. The performance of the classifier is then evaluated based on the accuracy in classifying test data points.

5.5.2 Machine Learning Algorithms

In most scenarios, multiple learning models are experimented and classification evaluation metrics are further compared to pick the classifier that performs best and meanwhile satisfies the time, complexity, and cost constraints. However, there always exist some insights about the problem in hand that can help guide the designer to exclude or include different algorithms. In our application, we only address the pool of learning algorithms that, to the best of our knowledge, we think is widely used across ML applications.

The selection of learning algorithm is inspired by Occam's Razor principle which states that, out of all possible models that provide similar results, the simplest one should be selected as the final model. Thus the criteria by which we choose the adopted learning algorithm is based on its simplicity and speed as compared to other algorithms.

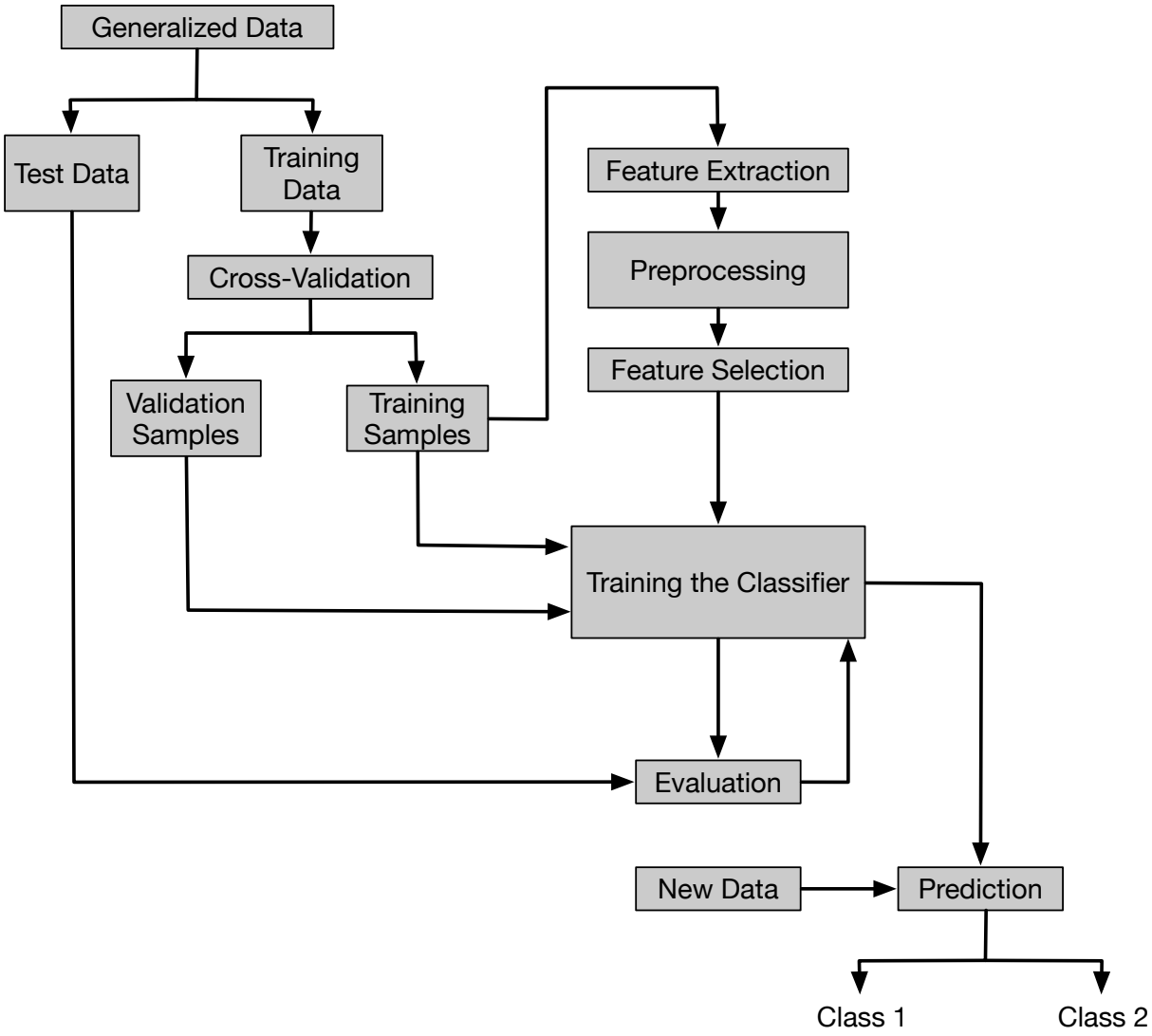


Figure 5.2: Training the classifier

5.5.3 Evaluation Metrics

As any other binary classification problem, there exist two possible kinds of classification errors. In a False Positive (F_P) decision, a negative class is classified as positive while in a False Negative (F_N) decision, a positive class is classified as negative. A positive instance that is correctly classified is a True Positive (T_P), while a negative instance that is correctly classified is a True Negative (T_N).

Accuracy (A), the most widely used evaluation metric, is the ratio of points correctly classified.

$$A = \frac{T_P + T_N}{T_P + F_N + T_N + F_P},$$

Recall (R) is the percentage of positives classified and Precision (P), the degree to which the classified positives are indeed positives. Recall and Precision are given as,

$$R = \frac{T_P}{T_P + F_N},$$

and

$$P = \frac{T_P}{T_P + F_P}.$$

5.6 The Utility Loss Metric

We measure the data utility based on the accuracy of the trained ML model. As shown in Algorithm 3, we first train a model using the original data then we re-train it using the generalized data. Comparing the classification accuracy in both scenarios gives an intuitive quantification of the utility loss resulting from privacy constrained data publishing.

The maximum data utility is achieved when a model is trained on the original dataset

without any generalization. The least data utility corresponds to a model that is trained using the most generalized dataset. Based on these bounds, the data utility loss is defined as follows.

Definition 24 (**Utility Loss $\mathscr{L}_{U_x}$**)**.** For a machine learning model M trained on published, generalized, dataset D_x corresponding to a privacy constraint ε_x, the utility loss, that is the cost of privacy, is defined as the drop in the model's accuracy A_x from a model trained on the original dataset D with an accuracy A_{max},

$$\mathscr{L}_{U_x} = \frac{A_{max} - A_x}{A_{max}}, \tag{5.3}$$

where A_{max} is the model's accuracy when trained on the original dataset, while A_x is the model's accuracy when trained using the generalized dataset D_x. The utility loss $\mathscr{L}_{U_x}$ is simply 0 when the model is trained using the original dataset which results in an accuracy A_{max}. The utility loss metric can be exploited to describe a publishing model's tolerance as follows.

Definition 25 (λ-**Utility Loss Tolerance**)**.** A data publishing model is said to be λ-utility-loss-tolerant if it tolerates up to λ utility loss for a machine learning model M trained on a published dataset D_x. That is, for any published dataset D_x derived from the original dataset D, the utility loss $\mathscr{L}_{U_x} < \lambda$.

The proposed metrics are intended to serve as foundation for utility-privacy tradeoff management in systems that conduct machine learning algorithms on the published data. Data publishers can use the proposed model to express the tradeoff as follows.

Definition 26 (λ-(ε, α)-**Utility-Privacy Tradeoff**)**.** A publishing model is said to have

Algorithm 3 The proposed iterative learning algorithm

For a given machine learning model M:

- Train M using the original data D
- Measure the model's accuracy A_{max}
- Generalize data based on a privacy loss constraint ε_x
- Train M using the generalized data D_x
- Measure the model's accuracy A_x
- Compute the data utility loss $\mathscr{L}_{U_x}$

λ-(ε, α)-utility-privacy tradeoff if the distribution and entropy privacy losses are constrained to ε and α respectively, and the utility loss tolerance is constrained to λ.

5.7 Privacy Preserving ATO Detection

In this section we introduce a privacy preserving machine-learning-driven classifier, designed to detect Account Takeover (ATO), a class of the most sensitive attacks in the field of email security. Specifically, we apply the proposed data publishing framework on a sample of the Microsoft Azure Active Directory data for employees in organizations that adopt Barracuda Networks' Sentinel as an advanced threat protection solution. The provided detailed implementation aims to help understand, analyze, and evaluate the capabilities of the proposed privacy framework.

We first provide a brief explanation of the ATO attacks and then introduce the proposed algorithm to utilize data from Barracuda's wide customer base in building an AI model that can tackle such attacks. We then describe the dataset and define its interesting attributes then identify which attributes can be considered personally identifying and which can be

considered quasi-identifiers or sensitive attributes. These attributes are then used to build the volumetric features that will be used to train the machine learning model which detects ATO attacks. To elaborate the idea behind the proposed framework, we then generalize the dataset, by generalizing the quasi-identifiers, and compute the utility and privacy losses in multiple scenarios to show how data owner and users can manage the utility-privacy tradeoff based on the proposed framework.

5.7.1 Account Takeover Attacks

Account Takeover (ATO) attack is one of the most sensitive attacks in the world of email security. According to [91], identity fraud hit all time high with 16.7 Million U.S. victims in 2017 only. During this year, businesses reported around $5.1 billion of losses due to ATO attacks. In an ATO, attackers get illegitimate access to users' accounts through a malicious login. Detection of compromised accounts and stopping the attackers before data is exfiltrated, destroyed, or the account is used in any nefarious actions is the ultimate goal for ATO detectors. Examining a sample of compromised accounts from multiple data sources, one prior study of account takeover discovered that attackers mostly use these accounts to send targeted phishing emails to the account's contacts [92]. Since these phishing emails are sent from legit contacts' accounts, they are extremely hard to detect using the existing phishing detectors at the recipients' side. This renders the phishing attacks undetected and results can be very costly.

An ATO can be detected through monitoring organizations' internal email traffic and classifying lateral phishing, that is malicious emails sent from one employee to another within the same organization [93]. As proposed in [94], ATO can also be detected using statistical behavior features extracted from graph topology including, success out-degree proportion,

reverse page-rank, recipient clustering coefficient and legitimate recipient proportion. While those two approaches are very effective, they are solely based on the email traffic. In other words, if the attacker compromises an account and only spies on the compromised users, then those detection classifiers will never be triggered. Another method of detection, that can handle this case, is monitoring the login attempts for employees and building models to find suspicious activity such as irregular login locations, times, devices, or browsers. By extracting timely sign-in data from Office365 logs, Sentinel builds multiple models that detect suspicious logins and alerts system admins accordingly.

Our work introduces a new class of ATO detection using login data. The proposed approach relies on assessing the fraudulence confidence level of login IP-addresses and user-agents to classify each login. The IP-address, and the User Agent, a field that provides information about the device and browser used in the login, are the interesting fields for this class. As we will show later, these two attributes make it possible to ascertain if a login instance is suspicious. This class works best with detecting a kind of attack where password-spraying login attempts are performed to compromise accounts. Password-spraying attacks are performed by using a large number of usernames and combining them with a single password. Unlike brute-force attacks, where one username is used with many password variations, password-spraying attacks avoid account lock-out because they look like isolated failed logins. Attackers exploit the large credential dumps to find common variations of usernames and passwords.

5.7.2 Data Exploration

Office365 is becoming a repository of valuable organizational information. The proposed model exploits a sample of the Microsoft Office365 Azure Active Directory data for employ-

Table 5.1: Description of Microsoft azure active directory dataset

	Attribute	**Type**
1	ClientIP	QID
2	OrganizationId	QID
3	Country	QID
4	ExtendedProperties	QID
5	Operation	QID
6	CreationTime	QID
7	Fraudulence	Sensitive

ees in organizations that adopt Barracuda Networks' Sentinel as advanced threat protection solution. The dataset records express all login attempts to office365 accounts. The dataset comprises a total of 20 fields that provide information about company, user, date, device, browser, authentication, and login IP-address. As shown in Table 5.1, after removing Personally Identifying Information (PII) such as name and email, we utilize only 7 attributes, 6 of which form the set of possible quasi-identifiers and the 7^{th} is the sensitive attribute. Namely, quasi-identifiers are the ClientIP which is the IP-address that an employee logged-in from, the OrganizationId that the employee belongs to, the Country of the login, CreationTime of the login, the Operation result that states if the login was successful or not, and finally the ExtendedProperties that provide information about the browser, device, and authentication used in the login. The Fraudulence is an added field to represent the label that decides the legitimacy of a login and this is the sensitive attribute. All attributes are defined in [95].

The raw data is processed to yield an augmented dataset that provides statistical features for each logging IP-address. The augmented dataset serves as the original dataset of publishing interest. We study the effect of publishing this dataset, and different generalized versions of it, on the data privacy as well as the data utility losses.

5.7.3 Feature Extraction

The proposed model relies on a set of features to detect the reputation of the IP-addresses. This reputation is then utilized to classify users' logins. We augment the raw-data's attributes to extract a set of features that would be useful in machine learning model's training. This set relies mainly on the statistical data that we extract from logins across the wide customer range. $IPBadLogins$ and $IPGoodLogins$ are the total counts of failed and successful logins from the IP-address respectively. $IPBadUsers$ and $IPGoodUsers$ are the total counts of distinct users with failed and successful login attempts from the IP-address respectively. $IPBadOrgs$ and $IPGoodOrgs$ are the total counts of distinct organizations with failed and successful logins from an IP-address respectively.

We parse the feature $ExtendedProperties$ to extract the user agent. The user agent is then processed to generate the last feature, $UserAgentFlag$. The $UserAgentFlag$ is a binary feature that takes the value 1 if the user agent is suspicious and 0 otherwise. Exploring the data, we found that there are some user agents that are usually connected with brute force ATO attacks. An example of such user agents is the CBAInPROD. While there isn't a confirmed explanation of the nature of such user agent, it is generally believed to be connected with logins that utilize IMAP (Internet Message Access Protocol) to perform password-spraying attacks. IMAP is a legacy authentication protocol that makes it possible for an account to be accessed from multiple devices. The protocol that does not support MFA is often used by desktop email clients to retrieve emails from the email servers. It is also worth to mention that the user-agent field can be easily manipulated by the attackers. This implies that, in some cases, this class of attacks will not necessarily come from a CBAInPROD user-agent.

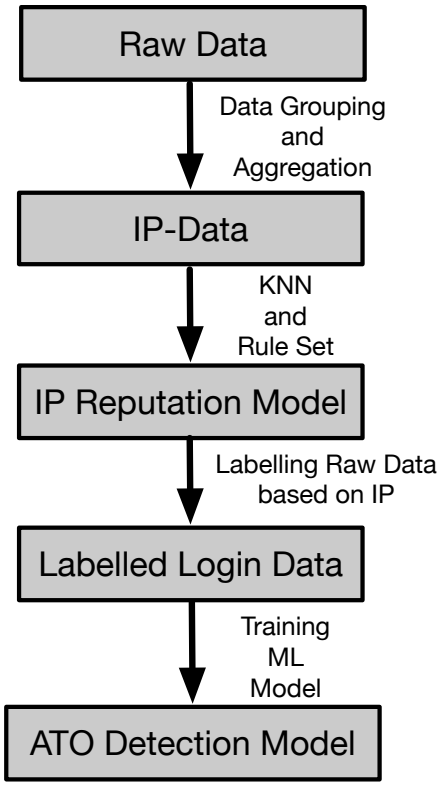


Figure 5.3: The proposed ATO detection Model

5.7.4 Training on Original and Generalized Data

Datasets and Labeling Three types of datasets are generated in our work. First is the raw-data, where we exploit a sample of Barracuda's entire raw login data over the period of 3 months. This sample contains, approximately, 700M data points. A sample of the raw-data's structure is shown in Table 5.2. The raw-data is used to generate the second type of datasets, IP-data. The IP-data is generated by grouping the raw-data by ClientIP (also named IP for the sake of consistency) and aggregating the statistical features described in the previous subsection. The resulting dataset contains, approximately, 8M data points, where each data point is a distinct IP-address and the corresponding columns are the statistical data connected with this IP. A sample of the IP-data's structure is shown in Table 5.3. Now that we generated a dataset that reflects the behaviour by which each IP-address is used, we

Table 5.2: Sample raw-data

UserId	OrganizationId	CreationTime	UserAgent	Operation	IP	Country
user@domain.com	ABCDXX	2020-03-26T12:00:00Z	Mac OS X - Chrome	UserLoggedIn	x:x:x:x	US

can start finding criteria by which we can build an IP reputation Model.

We adopt two classes of IP reputation; Bad and Good. To learn how to decide whether an IP should be reputed as Bad or Good, we collect few random samples of known IP-addresses that are connected with ATO incidents that exploit the password spraying behavior. Statistical feature values for these IP-addresses are then collected from the raw-data. Based on these samples and their collected feature values, we train a K-Nearest Neighbor (KNN) model to capture more data points and thus more Bad reputation IP-addresses. Another KNN model is also trained on a random sample of known legit logins from IP-addresses that are known to be safe. The collected data for both classes aided in setting the criteria, represented in a rule set, by which we split the 8M IP-dataset into the two classes. Approximately, 16K IP-addresses were marked as Bad reputation IPs, while the remaining IPs where marked as Good.

The IP-data with the corresponding reputation is then used to form our third dataset; the labeled dataset. To generate this dataset, we collect random samples of logins that took place from each set of Bad and Good reputation IP-addresses. A login is labelled as Fraudulent if it took place from a Bad reputation IP-address, otherwise, it is labelled as a Legit login. The resulting labelled data contains 1M Legitimate, and 0.13M Fraudulent logins, respectively. The users in the labelled dataset are spread across 1886 companies. This dataset is used to train the proposed ATO-detection ML Model. Figure 5.3, describes the transition and means by which the labelled data is generated from the raw login data passing through the generated IP-data.

Table 5.3: Sample IP-data

IP	$IPBadLogins$	$IPGoodLogins$	$IPBadUsers$	$IPGoodUsers$	$IPBadOrgs$	$IPGoodOrgs$	$UserAgentFlag$

Model Training and Classification Labeled Data is split into training and test with a ratio 0.75 and 0.25 respectively, where 850854 data points are used in training the model while 283619 are test points. A Random Forest [96] classifier is trained using the labelled data. Most machine learning models, including Random Forest, require the user to set various hyper-parameters that govern the model's training process. To determine the optimal set of hyper-parameters for our classifier, we followed machine learning best practices by conducting a three-fold cross-validation grid search over all combinations of the hyper-parameters listed below [97].

- Number of trees: 10–200, in steps of 5 (i.e., 15, 20, 25, . . . , 175, 200)
- Maximum tree depth: 10–100, in steps of 5 (i.e., 15, 20, 25, . . . , 95, 100)
- Minimum leaf size: 1, 2, 4, 8, 16

Our model used a Random Forest model with 75 trees, a maximum depth of 15, a minimum leaf size of 8 elements. Once the classifier has been trained, given a new login, the features are extracted by joining the login IP with the IP-data. The new login and the extracted features are then fed into the trained classifier which then outputs the decision.

5.7.5 Utility-Privacy Tradeoff Management

We aim at computing the privacy losses resulting from publishing the generalized dataset as well as the drop in the ML model's accuracy as a result of the dataset generalization. After generalization, a dataset is split into Q classes. Each class brings on its own privacy

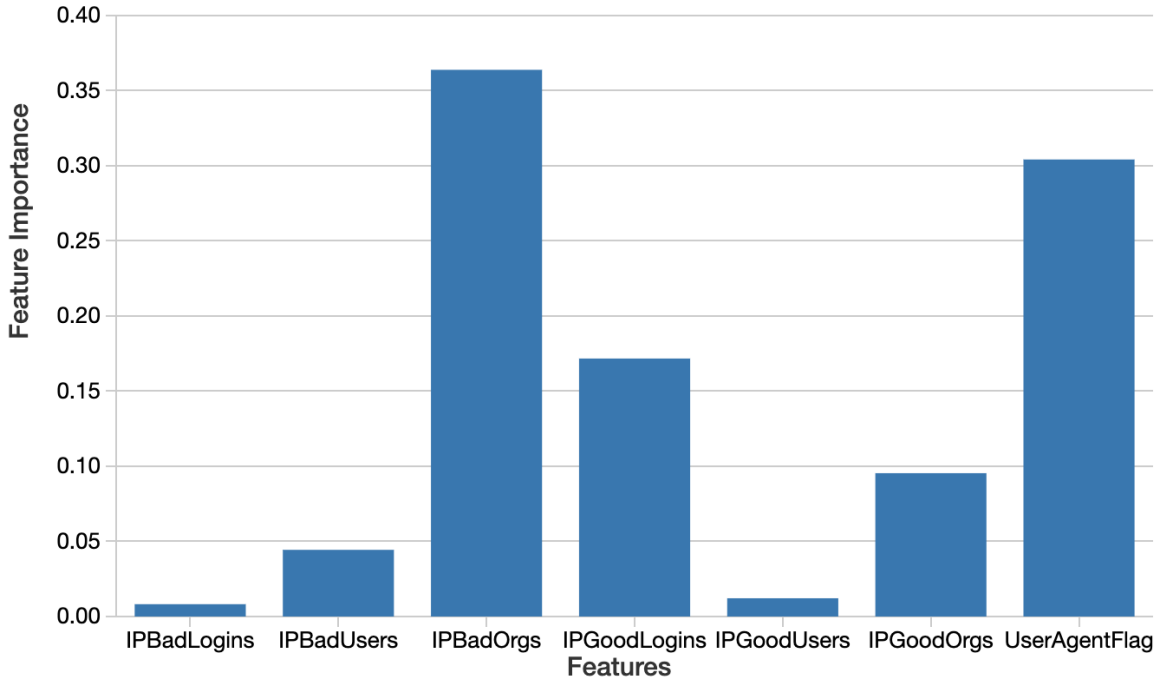


Figure 5.4: Feature importance of original-data trained model

losses. Collectively, these losses will contribute to the total privacy loss of the published dataset. For example, if the distribution of the sensitive attribute in the original dataset is (0.88,0.12). A published class with sensitive attribute distribution (0.1,0.9) would reveal an accountable information about users that fall in this class. If this class involves all logins from some country X, then knowing that a user logs in from this country would drastically affect the belief about the login's fraudulence.

We study two variants of generalization. First is using the IP sub-net rather than the IP-address. Second is using the Country and omitting the IP from the dataset. Table 5.4 expresses the features used in each variant. Figures 5.4, 5.5, and 5.6 show the feature importance in 3 different models trained using the original dataset, sub-net generalized dataset, and the country generalized dataset. As shown in Table 5.6, after generalization, the original dataset is split into $Q = 21368$ classes in the sub-net dataset and $Q = 272$ classes in the country dataset.

Table 5.4: Feature sets from different generalized datasets

Original Features	**Sub-net Features**	**Country Features**
$IPBadLogins$	$SubnetBadLogins$	$CountryBadLogins$
$IPGoodLogins$	$SubneGoodLoginst$	$CountryGoodLogins$
$IPBadUsers$	$SubnetBadUsers$	$CountryBadUsers$
$IPGoodUsers$	$SubnetGoodUsers$	$CountryGoodUsers$
$IPBadOrgs$	$SubnetBadOrgs$	$CountryBadOrgs$
$IPGoodOrgs$	$SubnetGoodOrgs$	$CountryGoodOrgs$
$UserAgentFlag$	$UserAgentFlag$	$UserAgentFlag$

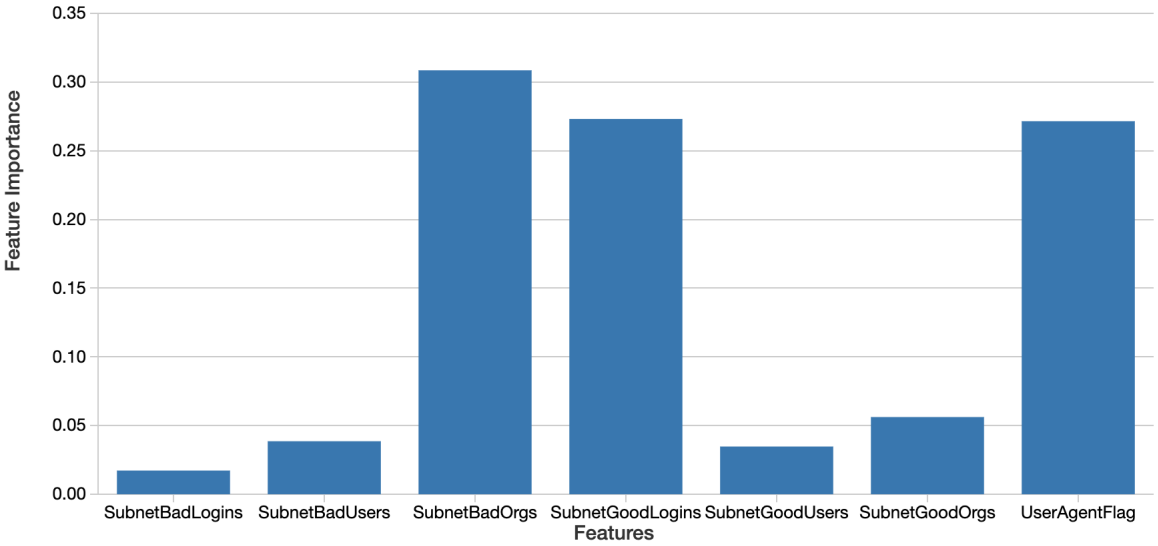


Figure 5.5: Feature importance of subnet-generalized-data trained model

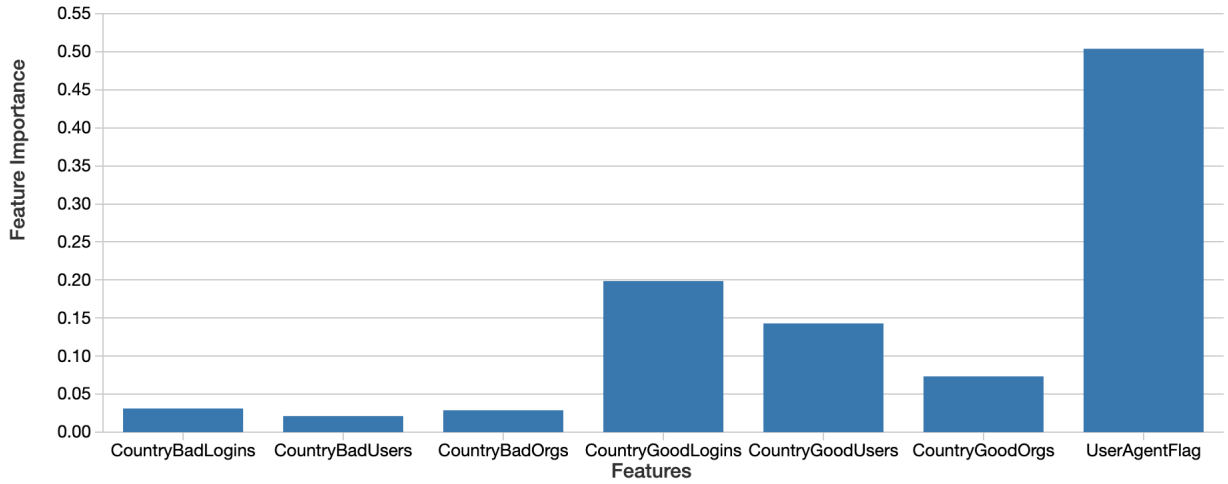


Figure 5.6: Feature importance of country-generalized-data trained model

Accuracy, precision, and recall are computed for the three variants and shown in Table 5.5. Privacy losses are computed according to Equations (5.1) and (5.2) while the loss in data utility is computed based on Equation (5.3). The results show that maximum, minimum, and average privacy losses maintain a consistently decreasing pattern as models are trained using more generalized datasets. However, as predicted, data utility loss shows an increasing pattern. Although this inverse relation between privacy and utility losses is expected, it is very interesting to notice the relative change in both metrics. According to Table 5.6, if the sub-net generalized dataset is published rather than the original dataset, average privacy loss drops from maximum value to (0.1338, 0.1661) at the expense of only an 0.045% drop in data-utility. Additionally, if the country generalized dataset is published instead of the sub-net generalized dataset, average privacy loss drops to (0.0333, 0.0224) at the expense of 0.46% drop in data utility. While in many scenarios data utility loss is intolerable, these low data utility loss values might very well be negligible in most of the applications. This answers one of our most crucial questions; why would a data publisher release datasets that would leak a considerable amount of private information, when they could publish generalized versions of these datasets in an attempt to leak less private information at an

Table 5.5: Model results of training on original and generalized datasets

	Original Data	**Sub-net Generalized**	**Country Generalized**
T_P	33581	33509	33256
T_N	249996	249960	249006
F_P	4	40	994
F_N	16	110	341
$Precision(\%)$	99.98%	99.88%	97.09%
$Recall$ (%)	99.95%	99.67%	98.98%
$Accuracy$ (%)	99.99%	99.94%	99.52%

Table 5.6: Tradeoff results when training on original vs. generalized datasets

	Original Data	**Sub-net Generalized**	**Country Generalized**
$\#ofClasses$	–	21368	272
$Max.PrivacyLoss$ (ε, α)	–	(0.87, 0.5293)	(0.87, 0.5293)
$MinPrivacyLoss$ (ε, α)	–	(7.79e-7, 3.436e-6)	(1.73e-7, 1.78e-7)
$AveragePrivacyLoss$ (ε, α)	–	(0.1338, 0.1661)	(0.0333, 0.0224)
$DataUtilityLoss(\%)$	0%	0.045%	0.46%

inconsiderable data utility loss?. The proposed framework exactly answers this question by providing quantifiable measures that are intended to aid the data publishers in making their decision.

Our PySpark Random Forest classifier provides a built-in estimate of each feature's relative importance [98], where each feature receives a score between 0.0–1.0 and the sum of all the scores adds up to 1.0. We notice that in case of training the model on the original dataset, Figure 5.4 shows that the most important feature is the count of $IPBadLogins$. As the dataset is more generalized, the classification tends to rely more on the $UserAgentFlag$ feature. This can be noticed by checking the gradual increase in this feature's importance

shown in Figures 5.4, 5.5 and 5.6. As shown in Table 5.5, this results in an increased false positive and false negative rates and therefore, a minimized accuracy, precision, and recall. While using a model that is trained on the country generalized dataset does not provide a significant data utility loss, the high false positive and false negative rates might not be accepted in an application that protects users' security. Rates such as 994 false positive alerts and 341 missed ATO attacks might not be what a global security product is looking to provide to its customers. Thus, using the proposed framework, a designer can easily quantify, understand, and communicate the tradeoff and select the appropriate levels of utility and privacy loss depending on the application's expected performance.

5.7.6 Ethics

In this work, our team, consisting of researchers from academia and a large security company, developed privacy preserving ATO detection techniques using a dataset of historical account logins and reported ATO incidents from 1886 organizations who are customers of Barracuda Networks. These organizations granted Barracuda permission to access their Office 365 employee mailboxes. Per Barracuda's policies, all fetched emails and login data are stored encrypted. Only authorized employees at Barracuda were allowed to access the data (under standard, strict access control policies). No personally identifying information or sensitive data was shared with any non-employee of Barracuda. Once Barracuda deployed a set of ATO detectors to production, any detected attacks were reported to customers in real time to prevent financial loss and harm.

5.8 Summary

A framework to address the utility-privacy tradeoff in machine learning applications was proposed. The framework provided a quantification of the data utility loss from a ML perspective as a result of applying privacy constraints in data publishing. The proposed work showed how a data owner is able to manage the utility-privacy tradeoff and gain deeper insights about the value of the released data as well as the potential privacy losses. To demonstrate how the proposed framework can be applied in real life applications, an employees email login dataset from a top cybersecurity company was utilized. The ATO attacks were introduced together with a proposed ML-based detection algorithm. Results showed that almost same detection accuracy levels can be achieved when using anonymized data instead of the original data. We also showed that, depending on the DO's and DU's requirements including application's expected performance, the tradeoff can be easily quantified, understood, and communicated using our proposed framework.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we introduced a comprehensive characterization and novel quantification methods of privacy to deal with the problem of privacy quantification in privacy-preserving data publishing. In order to consider the privacy loss of combined attributes, we presented data publishing as a multi-relational model. We re-defined the prior and posterior beliefs of the adversary. The proposed model and adversarial beliefs contribute to a more precise privacy characterization and quantification. Supported by insightful examples, we then showed that privacy could not be quantified based on a single metric. We proposed two different privacy loss metrics.

Based on these metrics, the privacy loss of any given PPDP technique could be evaluated. We introduced two data utility loss metrics. Using these metrics, we were able to practically address the utility-privacy tradeoff problem. We then propose a utility-boosting privacy-preserving data disclosure model that redefines the data utility based on the DU's perspective. Based on this model we incorporate our utility and privacy metrics to propose two versions of a privacy-preserving data disclosure protocol. The protocol sets rules for the negotiation between the DO and the DU in order to set a data disclosure deal. The proposed protocol inherently boosts the data utility from the DU's perspective with the satisfaction

of the DO's privacy constraint and monetary objectives. Our experiments demonstrate how we could gain a better judgment of existing techniques and help analyze their effectiveness in reaching privacy.

Furthermore, a framework to address the utility-privacy tradeoff in machine learning applications was proposed. The framework provided a quantification of the data utility loss from a ML perspective as a result of applying privacy constraints in data publishing. The proposed work showed how a data owner is able to manage the utility-privacy tradeoff and gain deeper insights about the value of the released data as well as the potential privacy losses.

To demonstrate how the proposed framework can be applied in real life applications, an employees email login dataset from a top cybersecurity company was utilized. The ATO attacks were introduced together with a proposed ML-based detection algorithm. Results showed that almost same detection accuracy levels can be achieved when using anonymized data instead of the original data. We also showed that, depending on the DO's and DU's requirements including application's expected performance, the tradeoff can be easily quantified, understood, and communicated using our proposed framework.

6.2 Future Work

There is a continuous interest in building privacy preserving data publishing models that satisfy the privacy constraints and meanwhile provide the maximum data utility. Our work opens doors to a wide range of research directions that raise some problems and interesting questions. In the following, we highlight some of those research directions that can build on the work presented in this thesis.

- *Sufficiency of the proposed metrics* We raise a question regarding the sufficiency of different utility and privacy metrics in terms of fully accounting to the losses. Meanwhile, we do not know how many metrics would be sufficient to quantify privacy and utility losses. However, we believe that any further proposed independent metrics that would contribute to reaching an optimum and provably sufficient set of measures, can be added to the proposed quantitative measurement framework.

- *Achieving an optimal data disclosure model* Another open problem is the optimization of the original data generalization as to achieve maximum privacy based on our proposed metrics. Typically, we believe that equivalence classes should be designed in such a way that keeps both the entropy loss and the distribution loss below a certain pre-determined level. This motivates us to think of a typical publishing scenario. Moreover, exploiting the proposed metrics, the utility-privacy tradeoff can be extensively researched as an optimization problem to reach a provably optimum data disclosure model.

- *Designing publishing models with adjustable class privacy* We also leave as an open problem for further research, optimization of the chosen set of quasi-identifiers with an objective of minimizing distribution and entropy losses within the published table or specific classes of higher privacy concerns.

- *Game theoretic approach to model the tradeoff based on our proposed metrics* Our formalization to the data privacy and data utility including the proposed metrics can be further used to provide a game theoretic approach to manage the utility-privacy tradeoff. In this model, data users are interested to run some machine learning algorithm on the provided data. The data user provides some tolerance threshold on the data utility

loss represented as the drop in the ML model's accuracy. The released data is useful only if the data utility loss does not go below the specified tolerance threshold. Data Collectors are willing to maintain the privacy constraints promised to data providers. Thus, the data collector and the data user are looking forward to find a generalized version of the dataset that satisfies both privacy and utility constraints. The model would be a sequential game with perfect and complete information. More specifically, both players know the provider's behavior model. The data user also knows the data collector's available actions and preferences. The game starts with an offer from the data user to the data collector. In the offer, the required value for privacy parameters and the price must be specified. The game is held until an equilibrium is found.

- *Collaborative privacy preserving ATO detection* The proposed privacy preserving ATO detection framework could be extended to design a collaborative privacy preserving ATO detection algorithm where multiple security companies are interested to share IP reputation datasets in order to get a better precision and recall. This algorithm should provide strong guarantees on privacy of each company's dataset and also guarantees on the collective usability of the aggregated data. It would also be interesting to provide a monetary framework to incentive different companies based on their contribution to the data utility.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] L. Sweeney, "Simple demographics often identify people uniquely," *Carnegie Mellon University, Data Privacy*, 2000. [Online]. Available: http://dataprivacylab.org/projects/identifiability/

[2] ——, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[3] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security & Privacy*, 2008, pp. 111–125.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, Mar. 2007. [Online]. Available: http://doi.acm.org/10.1145/1217299.1217302

[5] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE*, 2007, pp. 106–115.

[6] N. Li, W. Qardaji, D. S. Purdue, Y. Wu, and W. Yang, "Membership privacy: A unifying framework for privacy definitions," in *CCS*, Berlin, Germany, 2013.

[7] I. Wagner and D. Eckhoff, "Technical privacy metrics: a systematic survey," *CoRR*, vol. abs/1512.00327, 2015. [Online]. Available: http://arxiv.org/abs/1512.00327

[8] A. P. Singh and D. Parihar, "A review of privacy preserving data publishing technique," *International Journal of Emerging Research in Management &Technology*, vol. 2, 2013.

[9] Y. Xu, T. Ma, M. Tang, and W. Tian, "A survey of privacy preserving data publishing using generalization and suppression," *Applied Mathematics & Information Sciences*, vol. 8, 2014.

[10] W. Fang, X. Z. Wen, Y. Zheng, and M. Zhou, "A survey of big data security and privacy preserving," *IETE Technical Review*, vol. 34, no. 5, pp. 544–560, 2017. [Online]. Available: https://doi.org/10.1080/02564602.2016.1215269

[11] Y. H. Liu, Z. Tieying, J. Xiaolong, and C. Xueqi, "Personal privacy protection in the era of the big data," *J. Comput. Res. Dev.*, vol. 52, p. 1, 2015.

[12] Y. Rubner, C. Tomasi, L. J., and Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[13] S. A. Onashoga, B. A. Bamiro, A. T. Akinwale, and J. A. Oguntuase, "Kc-slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes," *Information Security Journal: A Global Perspective*, vol. 26, no. 3, pp. 121–135, 2017. [Online]. Available: https://doi.org/10.1080/19393555.2017.1319522

[14] T. S. Gal, Z. Chen, and A. Gangopadhyay, "A privacy protection model for patient data with multiple sensitive attributes," *International Journal of Information Security and Privacy*, vol. 2, p. 28, 2008.

[15] J. Han, F. Luo, J. Lu, and H. Peng, "Sloms: A privacy preserving data publishing methods for multiple sensitive attributes microdata," *Journal of Software*, vol. 8, p. 12, 2013.

[16] J. Li, R. C. Wong, A. W. Fu, and J. Pei, "Achieving k-anonymity by clustering in attribute hierarchical structures," *Data Warehousing and Knowledge Discovery*, vol. 4081, p. 405, 2006.

[17] Q. Liu, H. Shen, and Y. Sang, "Privacy-preserving data publishing for multiple numerical sensitive attributes," *Tsinghua Science and Technology*, vol. 20, p. 246, 2015.

[18] N. V. Mogre, G. Agarwal, and P. Patil, "Privacy preserving for high-dimensional data using anonymization technique," *International Journal of Advanced Research in Computer Science and Software Engineering Research*, vol. 3, 2013.

[19] A. K. Mohan, M. A. Phanindra, and M. K. Prasad, "Anonymization technique for data publishing using multiple sensitive attributes," *IJCST*, vol. 3, 2012.

[20] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 517–526. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557079

[21] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A framework for efficient data anonymization under privacy and accuracy constraints," *ACM Trans. Database Syst.*, vol. 34, no. 2, pp. 9:1–9:47, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1538909.1538911

[22] W. Liao, J. He, S. Zhu, C. Chen, and X. Guan, "On the tradeoff between data-privacy and utility for data publishing," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, 2018, pp. 779–786.

[23] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 3:1–3:36, Jan. 2014. [Online]. Available: http://doi.acm.org/10.1145/2514689

[24] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: ACM, 2014, pp. 1447–1458. [Online]. Available: http://doi.acm.org/10.1145/2588555.2588581

[25] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *Trans. Info. For. Sec.*, vol. 8, no. 6, pp. 838–852, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1109/TIFS.2013.2253320

[26] V. Rastogi, D. Suciu, and S. Hong, "The boundary between privacy and utility in data publishing," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, ser. VLDB '07. VLDB Endowment, 2007, pp. 531–542. [Online]. Available: http://dl.acm.org/citation.cfm?id=1325851.1325913

[27] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *CoRR*, vol. abs/1511.02381, 2015. [Online]. Available: http://arxiv.org/abs/1511.02381

[28] Y. Wang, Y. O. Basciftci, and P. Ishwar, "Privacy-utility tradeoffs under constrained data release mechanisms," *CoRR*, vol. abs/1710.09295, 2017. [Online]. Available: http://arxiv.org/abs/1710.09295

[29] F. du Pin Calmon, A. Makhdoumi, and M. Medard, "Fundamental limits of perfect privacy," in *IEEE International Symposium on Information Theory, ISIT 2015, Hong Kong, China, June 14-19, 2015*, 2015, pp. 1796–1800. [Online]. Available: https://doi.org/10.1109/ISIT.2015.7282765

[30] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 70–78, 08 2008.

[31] L. G. S. Giraldo and J. C. Príncipe, "Rate-distortion auto-encoders," *CoRR*, vol. abs/1312.7381, 2013. [Online]. Available: http://arxiv.org/abs/1312.7381

[32] K. Mivule and C. Turner, "A comparative analysis of data privacy and utility parameter adjustment, using machine learning classification as a gauge," *Procedia Computer Science*, vol. 20, no. Supplement C, pp. 414 – 419, 2013, complex Adaptive Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050913010958

[33] I. S. Rubinstein, "Big data: The end of privacy or a new beginning?" *N.Y.U. Public Law and Legal Theory Working Papers*, 2012.

[34] J. Janes, "As the big data beast fattens, will privacy and ethics get gobbled up?" *Am. Libraries*, 2012.

[35] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. on Knowl. and Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2009.190

[36] C. Dwork., "Differential privacy," *ICALP*, 2006.

[37] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 439–450. [Online]. Available: https://doi.org/10.1145/342009.335438

[38] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology — CRYPTO 2000*, M. Bellare, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 36–54.

[39] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 247–255. [Online]. Available: https://doi.org/10.1145/375551.375602

[40] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *SIGMOD Rec.*, vol. 33, no. 1, p. 50–57, Mar. 2004. [Online]. Available: https://doi.org/10.1145/974121.974131

[41] K. Lefevre, D. J. Dewitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," *In Proceedings of ACM SIGMOD*, pp. 49–60, 2005.

[42] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transaction Knowledge Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[43] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," *In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE)*, pp. 217–228, 2005.

[44] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," *In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE)*, pp. 205–216, 2005.

[45] ——, "Anonymizing classification data for privacy preservation," *IEEE Trans. Knowl. Data Engin.*, vol. 19, no. 5, pp. 711–725, 2007.

[46] V. S. Iyengar, "Transforming data to satisfy privacy constraints," *In Proceedings of the 8th ACM SIGKDD*, pp. 279–288, 2002.

[47] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," *Journal of Biomedical Informatics*, vol. 50, pp. 4 – 19, 2014, special Issue on Informatics Methods in Medical Privacy. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046414001403

[48] X. He, Y. Xiao, Y. Li, Q. Wang, W. Wang, and B. Shi, "Permutation anonymization: Improving anatomy for privacy preservation in data publication," in *New Frontiers in Applied Data Mining*, L. Cao, J. Z. Huang, J. Bailey, Y. S. Koh, and J. Luo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 111–123.

[49] X. Xiao and Y. Tao, "Personalized privacy preservation," *Proc. ACM SIGMOD*, pp. 229–240, 2006.

[50] N. Adam and J. Worthmann, "Security-control methods for statistical databases: A comparative study." *ACM Computing Surveys*, 1989.

[51] N. Victor, D. Lopez, and J. Abawajy, "Privacy models for big data: a survey," *International Journal of Big Data Intelligence*, vol. 3, p. 61, 01 2016.

[52] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A new approach to privacy preserving data publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, 09 2009.

[53] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *SIGMOD*, 2000.

[54] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy-preserving data mining," *PODS*, 2003.

[55] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the sulq framework," *PODS*, 2005.

[56] I. Dinur and K. Nissim, "Revealing information while preserving privacy," *PODS*, 2003.

[57] P. Liu, L. Wang, and X. Li, "Randomized perturbation for privacy-preserving social network data publishing," in *2017 IEEE International Conference on Big Knowledge (ICBK)*, 2017, pp. 208–213.

[58] J. Traub, Y. Yemini, and H. Wozniakowski, "The statistical security of a statistical database," *ACM Transactions on Database Systems*, 1984.

[59] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ICDE*, 2006.

[60] T. Truta and B. Vinay, "Privacy protection: p-sensitive k-anonymity property," *Proc. Int'l Workshop Privacy Data Management (ICDE Workshops)*, 2006.

[61] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, Jun. 2010. [Online]. Available: http://doi.acm.org/10.1145/1749603.1749605

[62] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," *In Proceedings of the ACM Conference on Management of Data (SIGMOD)*, pp. 437–486, 2008.

[63] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Model. Meth. Appl. Sci.*, vol. 1, 01 2007.

[64] M. Markatou, Y. Chen, G. Afendras, and B. G. Lindsay, *Statistical Distances and Their Role in Robustness*. Cham: Springer International Publishing, 2017, pp. 3–26. [Online]. Available: https://doi.org/10.1007/978-3-319-69416-0_1

[65] S. Vadhan, *The Complexity of Differential Privacy*. Cham: Springer International Publishing, 2017, pp. 347–450. [Online]. Available: https://doi.org/10.1007/978-3-319-57048-8_7

[66] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[67] J. Wicherts and M. Bakker, "Publish (your data) or (let the data) perish! why not publish your data too?" *Intelligence*, vol. 40, no. 2, pp. 73–76, 2012.

[68] M. Baker, "Is there a reproducibility crisis? a nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help," *Nature*, vol. 533, no. 7604, pp. 452–455 2016, 2016.

[69] D. Donoho, "50 years of data science," *Journal of Computational and Graphical Statistics*, vol. 26, pp. 745–766, 10 2017.

[70] M. H. Afifi, K. Zhou, and J. Ren, "Privacy characterization and quantification in data publishing," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018.

[71] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1175–1191. [Online]. Available: https://doi.org/10.1145/3133956.3133982

[72] E. Hesamifard, H. Takabi, and M. Ghasemi, "Deep neural networks classification over encrypted data," in *Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy*, ser. CODASPY '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 97–108. [Online]. Available: https://doi.org/10.1145/3292006.3300044

[73] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data," *IACR Cryptology ePrint Archive*, vol. 2014, p. 331, 2014.

[74] O. Ohrimenko, F. Schuster, C. Fournet, A. Mehta, S. Nowozin, K. Vaswani, and M. Costa, "Oblivious multi-party machine learning on trusted processors," in *Proceedings of the 25th USENIX Conference on Security Symposium*, ser. SEC'16. USA: USENIX Association, 2016, p. 619–636.

[75] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2016.

[76] M. Al-Rubaie and J. M. Chang, "Privacy Preserving Machine Learning: Threats and Solutions," *arXiv e-prints*, p. arXiv:1804.11238, Mar. 2018.

[77] B. Baron and M. Musolesi, "Interpretable machine learning for privacy-preserving pervasive systems," *IEEE Pervasive Computing*, vol. 19, no. 1, pp. 73–82, 2020.

[78] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "An information-theoretic approach to privacy," *CoRR*, vol. abs/1010.0226, 2010. [Online]. Available: http://arxiv.org/abs/1010.0226

[79] L. Xu, C. Jiang, J. Wang, Y. Ren, J. Yuan, and M. Guizani, "Game theoretic data privacy preservation: Equilibrium and pricing," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 7071–7076.

[80] M. H. Afifi, E. Zaghloul, T. Li, and J. Ren, "Ubnb-ppdp: Utility-boosting negotiation-based privacy preserving data publishing," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[81] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, p. 211–407, Aug. 2014. [Online]. Available: https://doi.org/10.1561/0400000042

[82] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015, pp. 909–910.

[83] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *CoRR*, vol. abs/1603.01887, 2016. [Online]. Available: http://arxiv.org/abs/1603.01887

[84] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 1895–1912. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman

[85] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*, 2016. [Online]. Available: http://dx.doi.org/10.1145/2976749.2978318

[86] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, "Membership privacy in microrna-based studies," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 319–330. [Online]. Available: https://doi.org/10.1145/2976749.2978355

[87] B. K. Beaulieu-Jones, W. Yuan, S. G. Finlayson, and Z. S. Wu, "Privacy-preserving distributed deep learning for clinical data," *CoRR*, vol. abs/1812.01484, 2018. [Online]. Available: http://arxiv.org/abs/1812.01484

[88] N. Hynes, R. Cheng, and D. Song, "Efficient deep learning on multi-source private data," *CoRR*, vol. abs/1807.06689, 2018. [Online]. Available: http://arxiv.org/abs/1807.06689

[89] N. Papernot, M. Abadi, Ú. Erlingsson, I. J. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *ArXiv*, vol. abs/1610.05755, 2017.

[90] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," *CoRR*, vol. abs/1904.02200, 2019. [Online]. Available: http://arxiv.org/abs/1904.02200

[91] "Identity fraud hits all time high with 16.7 million u.s. victims in 2017," 2018, https://www.javelinstrategy.com/press-release/identity-fraud-hits-all-time-high-167-million-us-victims-2017-according-new-javelin.

[92] E. Bursztein, B. Benko, D. Margolis, T. Pietraszek, A. Archer, A. Aquino, A. Pitsillidis, and S. Savage, "Handcrafted fraud and extortion: Manual account hijacking in the wild," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, ser. IMC '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 347–358. [Online]. Available: https://doi.org/10.1145/2663716.2663749

[93] G. Ho, A. Cidon, L. Gavish, M. Schweighauser, V. Paxson, S. Savage, G. M. Voelker, and D. Wagner, "Detecting and characterizing lateral phishing at scale," in *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 1273–1290. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/ho

[94] X. Hu, B. Li, Y. Zhang, C. Zhang, and H. Ma, "Detecting compromised email accounts from the perspective of graph topology," in *Proceedings of the 11th International Conference on Future Internet Technologies*, ser. CFI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 76–82. [Online]. Available: https://doi.org/10.1145/2935663.2935672

[95] "Detailed properties in the Office 365 audit log," https://docs.microsoft.com/en-us/microsoft-365/compliance/detailed-properties-in-the-office-365-audit-log.

[96] T. K. Ho, "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition - Volume 1*, ser. ICDAR '95. USA: IEEE Computer Society, 1995, p. 278.

[97] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. null, p. 281–305, Feb. 2012.

[98] "Apache spark. pyspark decisiontreeclassificationmodel v2.4.0." http://spark.apache.org/docs/2.4.0/api/python/pyspark.ml.html?highlight=featureimportance#pyspark.ml.classification.DecisionTreeClassificationModel.featureImportances.